information

MDPI

Article

Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers †

Yalemisew Abgaz [1,*], Amelie Dorn [2], Barbara Piringer [2], Eveline Wandl-Vogt [2] and Andy Way [1]

1 ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland; andy.way@adaptcentre.ie
2 Austrian Centre for Digital Humanities, Austrian Academy of Sciences, 1010 Vienna, Austria; amelie.dorn@oeaw.ac.at (A.D.); barbara.piringer@oeaw.ac.at (B.P.); eveline.wandl-vogt@oeaw.ac.at (E.W.-V.)
* Correspondence: yalemisew.abgaz@adaptcentre.ie; Tel.: +353-863-207-535

† This paper is an extended version of our conference paper: Abgaz, Y.; Dorn, A.; Piringer, B.; Wandl-Vogt, E.; Way, A. A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 21–29.

Received: 15 September 2018; Accepted: 21 November 2018; Published: 24 November 2018

Abstract: Extensive collections of data of linguistic, historical and socio-cultural importance are stored in libraries, museums and national archives with enormous potential to support research. However, a sizable portion of the data remains underutilised because of a lack of the required knowledge to model the data semantically and convert it into a format suitable for the semantic web. Although many institutions have produced digital versions of their collection, semantic enrichment, interlinking and exploration are still missing from digitised versions. In this paper, we present a model that provides structure and semantics to a non-standard linguistic and historical data collection on the example of the Bavarian dialects in Austria at the Austrian Academy of Sciences. We followed a semantic modelling approach that utilises the knowledge of domain experts and the corresponding schema produced during the data collection process. The model is used to enrich, interlink and publish the collection semantically. The dataset includes questionnaires and answers as well as supplementary information about the circumstances of the data collection (person, location, time, etc.). The semantic uplift is demonstrated by converting a subset of the collection to a Linked Open Data (LOD) format, where domain experts evaluated the model and the resulting dataset for its support of user queries.

Keywords: ontology; E-lexicography; semantic uplift; semantic modelling; questionnaires; linked data; linguistic linked open data

1. Introduction

Many organisations and individual citizens around the world have been collecting a significant amount of linguistic (lexicographic and lexical), historical, socio-cultural, demographic and geospatial data. Such data have been collected mainly using traditional data collection methods where data collectors distribute questionnaires and gather the responses manually. Among the various groups, lexicographers and linguists have been involved in collecting lexicographic and linguistic data to support research in the area. Since linguistic data covers various aspects of a society, such endeavours often have resulted in the collection of additional but related historical, socio-cultural, political and geospatial features. Research institutions, national museums or archival centres are in possession of such collections being treated as traditional resources of historical importance. Nowadays, museums, bibliographic centres, libraries and national archives adopt open-access policies [1] to support citizens' scientific inquiry [2–4].

However, the disclosure of such resources is challenging due to several shortcomings including a lack of standard documentation during the data collection and data conversion stages, a lack of

tools and techniques that support the potential future uses of the data, and absence of mechanisms for interlinking the various aspects of the data. The lack of complete documentation during the data collection phase poses a challenge in that vivid understanding of the semantics of the collected data becomes difficult. It is usually challenging to obtain a complete description of the data collection process and the description of the data elements. The challenge worsens when many of the concepts used at the start of the data collection evolve and take different meanings and shapes over time. Finding information about who, when, why and how a given dataset is collected is still a big challenge and becomes more difficult when little documentation is left behind. Consumers of such data, however, need to clearly understand the semantics in order to utilise it efficiently to support their scientific enquiry.

The second challenge of opening up such collections is the lack of machine-readable semantics. Most of the collected data depends on the available technology at the time the data was collected, and the tools and techniques that are available today were not then known. The current effort to make data traditionally collected machine-understandable requires proper semantics for it to be correctly interpreted and understood. Even if there are efforts to develop ontologies in different disciplines, the limited availability of ontologies to describe traditional collections is one of the obstacles for machines being able to discover and interpret legacy data.

The other problem that poses a challenge is interlinking the entities within the collection and across other similar collections. The absence of a standard vocabulary at the time of the data collection, the lack of consistent use of semantics or the absence of schema mapping between different versions are all challenges for interlinking entities. A schema definition or a data dictionary facilitates the interoperability and interlinking of the data. However, it requires further mapping from the schema to a standard vocabulary. Interlinking such historical collections using existing LOD techniques requires a deep understanding of the structure and the semantics of the collection in addition to the requirements of the knowledge of the domain [5].

Since there are only a few methods and techniques available to address all these challenges at the current time, it is difficult and time-consuming to open up such collections to researchers and citizen scientists. However, combinations of various techniques are available to reduce the required effort. Among these techniques, digitisation has played a significant role in processing the data and making it available in a digital format by scanning images, processing the texts using Optical Character Recognition (OCR) and manual transcription of the original data.

The Austrian Academy of Sciences digitised, in part, a collection of approximately 3.6 million paper slips and made them available in various formats including Tübinger System von Textverarbeitungs-Programmen (TUSTEP) files, Text Encoding Initiative/Extensible Markup Language (TEI/XML), as well as a MySQL database. Semantic modelling provides techniques to capture and model the semantics, and a mapping from non-RDF formats to LOD format is available to support such an endeavour. In this paper, we focus on the linguistic and historical data collection of Bavarian Dialects in Austria (DBÖ/dbo@ema) that covers data which were collected during the last century (1911–1998) and refers to a non-standard language of the beginnings of the German language up to the recent days. We propose to make available the collection of DBÖ/dbo@ema using an LOD approach. This research includes analysing the collection, proposing a semantic model for the collection, and the modelling of the core entities including questionnaires, questions, answers (lemmas, descriptions, pronunciations, illustrations), authors, collectors, respondents, and geographic locations. The questionnaires and questions are the essential parts of the entire collection as they serve as a semantic entry point to access the answers. 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 [6]. The questionnaires of DBÖ/dbo@ema which were created to collect the data [7] are used as a case study to demonstrate the process.

Our approach benefits from state-of-the-art LOD platforms [8] to support a more productive, enhanced and standard means of accessing the data for both human and machine agents by supporting

semantic browsing and SPARQL queries. It also allows the use of dereferenceable International Resource Identifiers (IRIs) to uniquely identify the resources and support their consistent interpretation using the semantic model supported by our ontology. The main contributions of this paper include:

1. Providing a semantic model for generic and domain-specific traditional data collections and analysis together with an ontology that provides the required semantics to interpret the content consistently.
2. Providing a semantic mapping to uplift the existing data to an LOD platform. We provide an R2RML mapping which will be used to transform the collection to an LOD following the W3C recommendations.
3. Providing an implementation and validation of the proposed approach that supports user requirements. To support this, we use common navigation paths that are extracted from the daily information requirements of existing users.
4. Additionally, capturing methods of integrating domain experts in the semantic modelling process and improved handling of changes during the semantic modelling and uplifting process.

In this paper, we demonstrate our approach using the subset of the DBÖ/dbo@ema which is available in the MySQL database developed in the project dbo@ema 2007-2010. This data includes 720 questionnaires, 24,382 questions, 11,157 individuals and organisations, 65,839 paper slips, 98,272 answers, 8,218 multimedia files and 16,839 sources. The resulting LOD data includes more than 2.8 million triples organised into eight named graphs. Although our approach covers the modelling of several entities including lemma, sources and multimedia, for the sake of brevity we will provide a detailed discussion of selected entities throughout the paper.

The remainder of the paper is structured as follows: Section 2 introduces DBÖ/dbo@ema, and the collection and digitisation processes. Section 3 provides details of the approach including the user requirements (Section 3.1), how the domain analysis process is carried out (Section 3.2), and the schema analysis process (Section 3.3). The semantic modelling of the entities and the ontology creation process are discussed in Section 4. The process of converting the data to an LOD using the resulting ontology and R2RML mapping is presented in Section 5. Section 6 discusses the implementation and validation process. We presented a comparison of our work with related research in Section 7, and we conclude, along with avenues for future work in Section 8.

2. Background

The study of linguistics in a historical context aims at understanding the use of language and its constructs in a society over time [9]. It focuses on collecting linguistic and lexicographic data that represent the language within a specific geographical location where the language of interest is used. Despite its focus on linguistic and lexicographic data collection, linguistic research is a discipline interwoven with the historical, social and cultural structure of the target society. Accordingly, the linguistic data collection methods usually go beyond collecting words and meanings of words and include the cultural context where the language is used: the history, demography and political aspects of the society. The data collected primarily includes linguistic features such as the naming of things, the meanings of words, phrases and sentences, the morphology, phonology and syntax of the language, and sometimes detailed descriptions of contexts and cultural backgrounds related to the use of specific words. The process of linguistic data collection is not restricted to simple words and their meanings, but also covers the culture of societies. Culture is expressed through the use of language [10], and even a single word can represent various meanings in different cultures and dialects.

Linguistic data collected over an extended period passes through various alterations to fit the changing requirements in time. Original data collection methods need updates; meanings of concepts may change through time and could denote a different meaning, and lots of the interpretation of the data will be changed due to a continuous change of personnel involved in the preparation of data collection tools. The continuous change in data collectors, respondents and data entry clerks leads to a considerable inconsistency in the interpretation of the entities, their attributes and classification of

the data in various categories which complicates the semantic enrichment process. To address these problems, in recent years, archival institutes, libraries, linguists and computer scientist have come together to set standards and semantic models to represent language resources [11–14].

The DBÖ/dbo@ema is a historical non-standard language resource which was originally collected under the Habsburg monarchy in paper slip format with the aim of documenting the Bavarian Dialect and rural life in Austria [7,15]. 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 [16] spoken by the locals, the data also 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 much more [15].

In response to the questionnaires (Figure 1a) distributed over the span of the project, close to 3.6 million individual answers noted on paper slips (Figure 1b,c) 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.

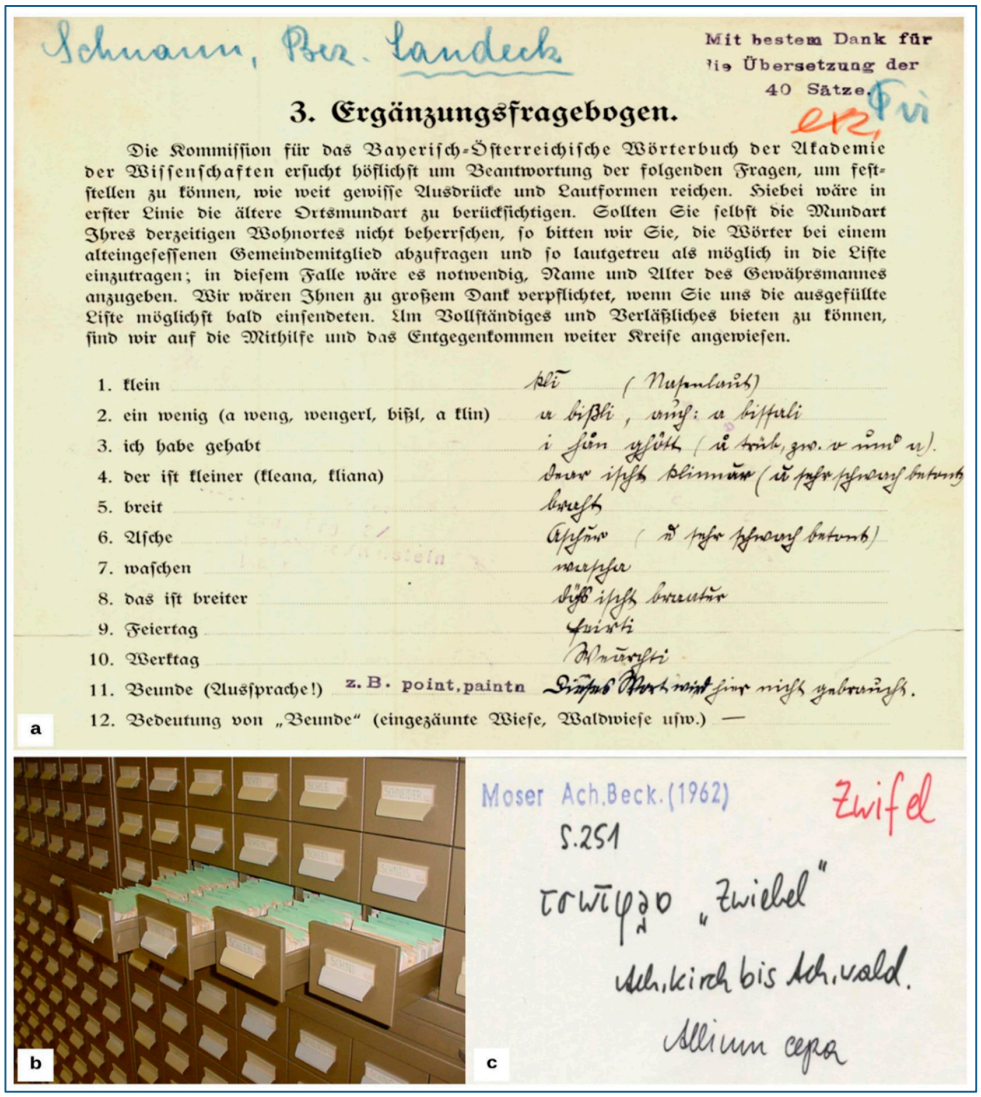

Figure 1. (**a**) Sample questionnaire; (**b**) the collection; (**c**) an individual paper slip.

In addition to the primary data, the entire collection also includes biographies of individual collectors and contributors. Several individuals who had various functions in the project had participated 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 during data collection and the digitisation process in later years is stored in a specific database (Personendatenbank [person database]). Persons and their background are thus key features of the data that offer additional perspectives for the exploration and the systematic opening of the collection. The dataset also holds 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 may include detailed information about the date and 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); to date, five volumes (A–E, including P and T) have been published [17]. Today, about three-quarters of the collected paper slips are available in digital format following several stages of digitisation, including scanned copies of the paper slips, questionnaires and textual representations of the paper slips in TUSTEP [18], MySQL [18] and TEI/XML [19]. This is an ongoing effort to make the data accessible and available for detailed analysis, including the use of semantic web technologies to make the data suitable for semantic publishing in the LOD platform.

3. The Approach and Development of a Semantic Model

Semantic publishing of traditional data using LOD platforms has become a focus for digital humanities research [20–22]. Semantic publishing involves the analysis and representation of the domain knowledge using the appropriate semantics mainly employing an ontology [23]. Whenever there is a suitable ontology that represents the domain knowledge adequately, this step becomes less relevant. However, for domains that do not have well-established semantics, this step is crucial in understanding and representing the core entities of the domain and their relationship. Although there are standard and well-established models to represent linguistic resources [11,24,25], there is a dearth of semantic models to capture, represent and link the data collection process with the actual collected data. Our focus in this project is not restricted to the resulting answers, but also the questionnaires, questions, authors, collectors and other relevant entities that are seen throughout the process.

3.1. User Requirements

The requirements outlined by users include the availability of a standard description of the core entities and an explicit interlinking between related entities in the collection. Users also want to see how the collection is linked to other similar collections elsewhere, and how intra-linking could be achieved to support a broader exploration. In other cases, users of the system could be independent machines that require one or more ontologies to support autonomous exploitation of the data through machine agents such as bots. To support this, we identified the following requirements that the semantic model should satisfy.

1. The model should formally represent the semantics of the core entities and their attributes as well as the relationships between these entities. This process includes:
    a. Identification of the major entities in the collections;
    b. Identification of useful and relevant attributes of the entities; and
    c. Identification of the major relationships that link those entities.
2. The model should be suitable to annotate the existing content semantically. The semantic uplift process should be able to generate LOD and be amenable to future changes and updates.

3. The model should support a structured query to allow users to construct queries based on their information requirements. It should further allow computer agents to access the data via APIs.
4. The model should reuse existing ontologies and vocabularies to supply rich semantics and interlinking.
5. It is preferable to provide multilingual support in English and German languages (with possible extension to other languages) with names of entities and their description appearing in both languages to support a wide range of users.

3.2. Modelling the Domain

Domain analysis is one of the primary inputs to the semantic modelling process. It provides fundamental knowledge about what concepts and relations the domain captures, and how the domain experts represent, interpret and use them. Our approach examines primary and secondary sources of information, investigating original materials and interviewing users and maintainers of the collection. To support this, we involved the domain experts who are directly working on the collection and who amassed in-depth knowledge. We ran several workshops and face-to-face meetings to understand the domain, and accurately represent the entities and their semantics. We further investigated various published and unpublished sources that describe the data collection, digitisation and usage of the data for dictionary compilation.

We followed an approach proposed by Boyce and Pahl [26] and Noy and McGuinness [27] to structure the domain analysis. The approach outlines four steps—Purpose, Source, Domain and Scope—to understand the domain and identify and capture the core entities. 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 focus on German language and diverse nationalities in the late Austro-Hungarian monarchy [28]. The primary data is collected using questionnaires which are prepared by experts in the Austrian Academy of Sciences. The questionnaires were distributed, and the data collectors collected the relevant data. Depending on the type of questionnaire, they returned either small notepads with the recorded answers or the completed questionnaires to the central office. The answers continued to arrive several years after their distribution.

In some cases, the collectors themselves were the respondents, in other cases, the respondents were individuals or group of respondents. The domain of the collection is mainly linguistic, but also touches historical, cultural, and political aspects. Concerning the scope of the domain analysis, although we cover many of the aspects identified above, in this paper we limit ourselves to modelling of the core entities. In addition to these entities, we cover in less detail related entities such as multimedia associated with the above entities, sources and geographic locations.

3.2.1. Domain Analysis of Questionnaires

The questionnaires are the starting point for our exploration of the domain. The academy holds an unpublished book of almost all the questionnaires which serves as the primary source for analysing the structure, hierarchy and attributes of the original questionnaires. This resource is available in a printed format but was later converted to a digital format which we used to identify core entities that need to be modelled. In addition to the entities identified in Table 1 from the questionnaires, domain experts further highlight the relationships between these entities.

3.2.2. Domain Analysis of Paper Slips

Another primary source of information is the paper slip. There are 3.6 million paper slips that were collected and catalogued, containing information including the answers to the questions, the details of the collectors, the place and the date of the collection. Some slips may include additional notes attached to them by those who process the data and include citations to other sources of the data. Table 2 shows some of the core entities that are identified during the analysis phase.

Table 1. Book of questionnaires.

Book of Questionnaires			
Entity @en	**Description @en**	**Entity @de**	**Description @de**
Questionnaire	A questionnaire represents a set of questions that are related to each other. A questionnaire contains metadata such as questionnaire identifiers, titles, agents and publication-related information.	Fragebogen	Ein Fragebogen stellt eine Reihe von Fragen dar, die miteinander in Beziehung stehen. Ein Fragebogen enthält Metadaten wie Fragebogenbezeichner, Titel, Agenten und Informationen zur Veröffentlichung.
Authors	Authors are agents or persons who prepare the questionnaires and the questions contained in them.	Autoren	Autoren sind Agenten oder Personen, die die Fragebögen und die darin enthaltenen Fragen vorbereiten.
Collectors	Collectors are defined in friend of a friend (FOAF) ontology, and we will reuse the definition provided in FOAF agent classes.	Kollektor	Die Sammler sind in der FOAF-Ontologie definiert und wir werden die in den FOAF-Agenten-Klassen enthaltene Definition übernehmen.
Questions	A question represents an expression used to request information. A question can be asked in various forms and seeks different kinds of answers. Based on this, a question is further divided into subclasses.	Frage	Eine Frage ist eine Äußerung, die eine Antwort zur Beseitigung einer Wissens- oder Verständnislücke herausfordert. Eine Frage kann in verschiedenen Formen gestellt werden und sucht nach verschiedenen Arten von Antworten. Basierend darauf wird eine Frage weiter in Unterklassen unterteilt.
Topics	A topic represents the main subject of a questionnaire or a question. A questionnaire may focus on a general topic such as "Food" and a question may cover subtopics such as "Traditional Food".	Thema	Ein Thema ist das Hauptthema eines Fragebogens oder einer Frage. Ein Fragebogen kann sich auf ein allgemeines Thema wie "Essen" konzentrieren und eine Frage kann Unterthemen wie "Traditionelles Essen" abdecken.

Table 2. Catalogue of paper slips.

Catalogue of Paper Slips			
Entity @en	**Description @en**	**Entity @de**	**Description @de**
Paper Slip	A paper slip represents the information contained on individually printed paper slips. A paper slip contains original answers to the questions in the distributed questionnaires and could further contain additional comments.	Belegzettel	Ein Papierzettel repräsentiert die Informationen auf einzelnen gedruckten Papierbelegen. Ein Zettel enthält originale Antworten auf die Fragen in den verteilten Fragebögen und kann zusätzliche Kommentare des Sammlers oder des Bearbeiters enthalten.
Source	A source is anything that is used as a source of information. A source could be a person, a document or any other thing.	Quelle	Eine Quelle ist alles, was als Informationsquelle dient. Eine Quelle kann eine Person, ein Dokument oder eine andere Sache sein.
Lemma	A lemma is a word which is used as a headword in a dictionary. Lemma in our context refers to the headwords that are used in (WBÖ) and (DBÖ/dbo@ema).	Lemma	Ein Lemma ist ein Wort, das als Stichwort in einem Wörterbuch verwendet wird. Lemma bezieht sich in unserem Zusammenhang auf die Stichwörter, die im WBÖ und in der DBÖ/dbo@ema verwendet werden.
Answer	An answer represents a written, spoken or illustrated response to a question.	Antworten	Eine Antwort repräsentiert eine schriftliche, gesprochene oder illustrierte Antwort auf eine Frage.

3.3. Schema Analysis

To date, three different systems have been used to manage the collection. The first system is called TUSTEP (Figure 2) which is a piece of software used to store textual information. This system was used to store the textual description of the digitised paper slips. The second one is TEI/XML where the TUSTEP data is converted into a TEI/XML format (Figure 2). A sizeable portion of the digitised data is converted to TEI/XML format which represents the majority of the paper slip record in the collection. However, it does not include details of the questionnaires other than a link to identify which question is answered in the paper slip.

The major drawback of all three systems is that there is no well-established schema definition or data dictionary associated with the data. Thus, understanding the content of the fields, their values, and the relationships with other tables is complicated. Accordingly, domain experts were involved in understanding, describing and representing useful elements. With the help of these domain experts, we identified the attributes and relationships of the entities. Even if the data stored in all three systems is instrumental, it is not in the format the LOD community requires [29,30]. Thus, entities that need further data cleaning and treatment are identified and corrected to guarantee the delivery of good quality data during semantic publishing.

Figure 2. The data in TUSTEP (left) and TEI/XML (right) system.

To ensure the capturing of the correct semantics of the attributes of the core entities, we generate a spreadsheet that contains all the attributes together with their descriptions. This activity enables us to attach a clear description of the entities and allows all domain experts to update the descriptions of the entities. Through a continuous engagement with the domain experts, the descriptions are updated continuously. A stable version of the spreadsheet is used as an input during the ontology creation phase. Although we did this exercise for all entities, for brevity, we present the descriptions of the questions and questionnaires in Table 3.

Table 3. Description of attributes of entities.

Question		
attribute@en	**attribute@de**	**Description**
number	nummer	"Number" of the single question (but without their respective questionnaire numbers), compiled like it is listed in the book of questionnaires
originalQuestion	originalfrage	Question in the entire length, edited by linguists
shortQuestion	kurzfrage	Shortened question (limited to one line); thus, usually strongly edited; (originally) to be displayed after the question number in the entries of the TUSTEP/xml-files of the DBÖ/dbo@ema; with an indication of more text in the original, if available (asterisk at the end)
originalData	originaldaten	Question in the entire length, edited by linguists before 2007—text based on the TUSTEP database entry
Questionnaire		
number	nummer	Number of the questionnaire like it is indicated in the headings of the questionnaires
title	titel	Title, heading of the questionnaire like it is indicated in the questionnaires
keyword	schlagwoerter	Thematic keywords matching the topic of the questionnaire
yearOfPublication	erscheinungsJahr	The year, when the questionnaire was finally sent out to the collectors.
authorId	autorenId	The creator(s) of the questionnaire
originalData	originaldaten	A questionnaire in its entire length
note	anmerkung	Fields for possible notes; currently the label of the person who entered the (unedited)
release	freigabe	Release of entry (by the data scientists) for further processing by the linguists
checked	checked	This means: review and additional processing (e.g., adding the correct lemma) by linguists is completed
wordToolbar	wordleiste	Entry to be considered in the MS Word bar (which was established for compiling WBÖ entries ~2005-2007)
print	druck	Entry is checked and can be considered for printing
online	online	This entry is released online (it will be visible on the dbo@ema website)
published	publiziert	This entry is already processed for the printed version of the WBÖ

4. The Semantic Model: OLDCAN

This section discusses the process and design choice for creating the Ontology for Linguistic Data Collection and ANalysis (OLDCAN). The ontology is built using Web Ontology Language (OWL) specification following the ontological principles outlined in [27,31,32]. In this model, we provide the definition of the concepts, object properties and data properties in English and German. We also reuse other well-known ontologies including Friend of a Friend(FOAF) [33], Dublin Core (dc) [34], DBpedia Ontology (https://wiki.dbpedia.org/services-resources/ontology) and others. The ontology diagram is presented in Figure 3. The ontology is available with an `oldcan` namespace pointing to https://explorations4u.acdh.oeaw.ac.at/ontology/oldcan.

Before the development of the OLDCAN ontology, we explored potential ontologies that could describe the domain of interest. The search included different ontology repositories such as linked open vocabulary (LOV) repository, Schema.org and other specialised engines such as the Watson semantic web search engine. Most of the searched ontologies contain one or more of the core entities together with some properties. Some of these ontologies provide classes with the same name but with a different semantics which do not represent our domain-specific requirements. The reason we create entities of our own while continuing to consume equivalent entities such as `dct:creator` is that we decided to keep the legacy terminology used in the collection while keeping it linked to the existing standard vocabularies. This allows existing users to use the legacy system without facing problems related to the new vocabulary. The ontology keeps track of such entities using owl:equivalentClass and owl:equivalentProperty.

However, generic ontologies such as FOAF and dc are found to be suitable to represent entities such as Agents (Persons, Groups, Organisations), Collectors, Editors and Publications and their attributes such as name, date of birth and address. We included the use of such ontologies either to represent the original data or to provide additional metadata. Domain-specific ontologies of high interest such as Ontolex-Lemon [13] are also used to represent most of the linguistic features.

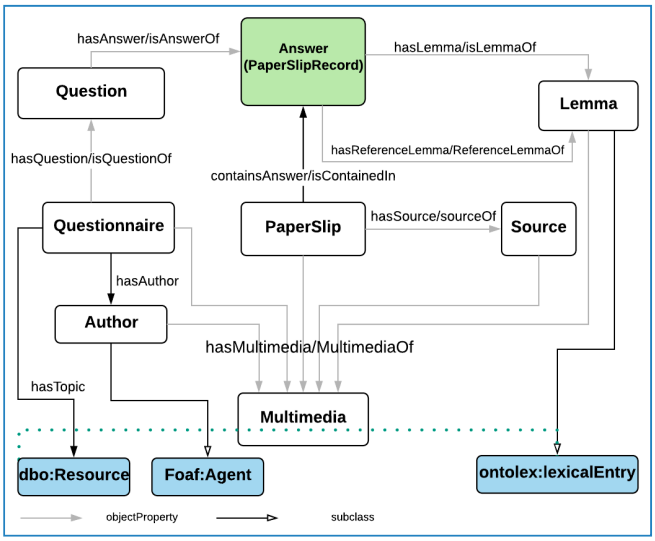

Figure 3. OLDCAN ontology diagram.

4.1. Concepts and Taxonomical Hierarchies

A detailed description of the core concepts and their taxonomical hierarchy and the design decisions is given below.

4.1.1. Questionnaire

The class oldcan:questionnaire represents all the questionnaires that are distributed during the data collection phase. This class is a top-level class which can represent any kind of questionnaire

independent of the domain. To meet the requirement of representing specific kinds of questionnaires kept in the collection, `oldcan:systematicQuestionnaire`, `oldcan:additionalQuestionnaire` and `oldcan:dialectographicquestionnaire` subclasses are created. These types of questionnaires have specific interpretations in the collection, e.g., `oldcan:additionalQuestionnaire` gives a complete sense only when it is interpreted with systematic questionnaires, as it is used to supplement the systematic questionnaires. A questionnaire may have one or more related questionnaires which are linked to it as a follow-up questionnaire or as a related questionnaire. This relation is captured using the `oldcan:hasRelatedQuestionnaire` object property with an inverse property of `oldcan:isRelatedQuestionnaireOf`. Each questionnaire in the collection represents a topic and contains several questions under it. Since there are several topics addressed by the questionnaire and since topic modelling is beyond the scope of this paper, we link the questionnaire topics with DBpedia resources (`dbo:resources`) with the `oldcan:hasTopic` property. This interlinking is done using a semi-automatic approach where the topics of the questionnaire are matched to DBpedia resources using DBpedia spotlight [35] and then manually evaluated and corrected by domain experts. The oldcan:questionnaire is linked to its oldcan:author via `oldcan:hasAuthor` property with an inverse property `oldcan:isAuthorOf`. For further enrichment, we provide the `dct:creator` property which is an equivalent property of `oldcan:hasAuthor`. The questionnaire model is presented in Figure 4.

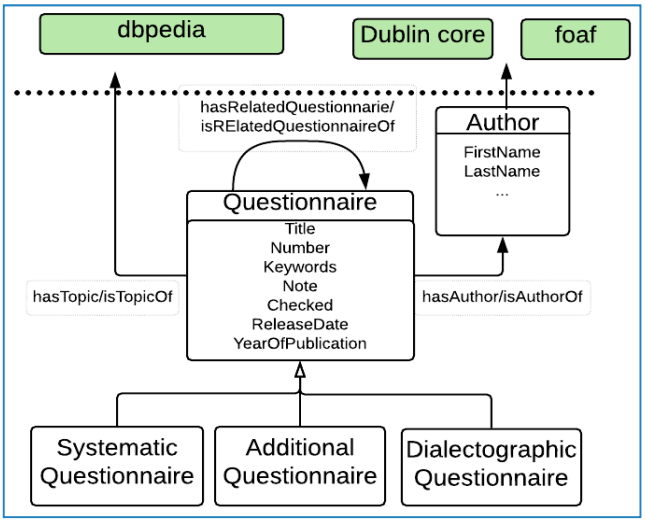

Figure 4. The semantic model for questionnaires.

4.1.2. Questions

The class oldcan:question (Figure 5) represents questions that are contained in the questionnaires. A question is a linguistic expression used to collect information or the request made using such expression [36]. The results of a question could be an answer to diverse types. Analysis carried out by the experts, users and ontology engineers identified 12 different types of questions. Conceptually, we categorise these 12 types into three conceptual levels: generic questions, linguistic questions and cultural questions. The generic question types are questions that apply to any domain. These classes include MultipleChoice, Dichotomous, Descriptive (open-ended), Ranking, Rating and Illustration questions. The linguistic level focuses on questions aimed at collecting specific linguistic features that distinguish them from the generic questions. These questions include Phonological, Morphological, Thesaurus, Syntactic, Onomasiological, Semasiological and Metaphorical questions. Cultural questions focus on questions that go beyond linguistic probes and encompass socio-cultural aspects. For instance, a question that asks how the naming of a given food is associated with a celebration is beyond linguistic questions that seek the naming of an entity. A description of each of the question types is given in Table 4.

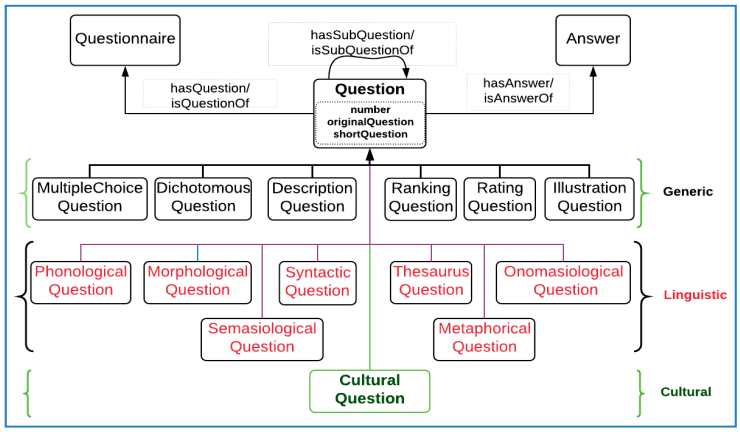

Figure 5. The semantic model of questions.

These three conceptual types become crucial in mapping the questions to their conceptual representation by the users of the system. For non-expert users, the questions can be either multiple choice, descriptive or any of the generic question types. However, for users with a linguistic and lexicographical background, the same question could be morphological, phonological or any of the lexical categories. For historians and cultural linguists, that same question could be viewed as a cultural question. To map this conceptual categorisation by different users, we used a flat taxonomy which allows a question to have one or more question types. For example, a semasiological question can also be a description (open-ended), multiple choice or rating questions. 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.

Table 4. Categorisation and descriptions of questions.

Level	Question Types	Description
Level-1: Generic	Multiple Choice	Asks for a selection of one item from a list of three or more potential answers.
	Dichotomous	Asks for a selection of answers from a binary option. It includes yes/no or agree/disagree types of answers to stated questions.
	Description	Asks for a written representation of a given entity, e.g., "What would be the function of x?".
	Ranking	Requires the respondent to compare entities and rank them in a given order.
	Rating	Asks the respondent to assign a rating (degree of excellence) to a given entity based on a predefined range
	Illustration	Asks for a pictorial or diagrammatic representation of a given entity, e.g., "What does x look like?".
Level-2: Linguistic	Phonological	Asks for the pronunciation or phonetic representation of words.
	Morphological	Asks about the structure and the formation of words and parts of words. Based on the structure, morphological questions can take various forms.
	Thesaurus	Asks for a list of words or expressions that are used as synonyms (sometimes, antonyms) or contrasts of a given entity.
	Syntactic	Demands the 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".
	Onomasiological	Asks for the name of a given entity, e.g., "how do you call x?" where x stands for an entity.
	Semasiological	Seeks the meaning of a given entity, e.g., "what does x mean?".
	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 apparent meaning.
Level-3: Cultural	Cultural	Asks for a belief of societies, procedures on how to prepare things, and how to play games, contents of cultural songs or poems used for celebrations.

It is commonly observed that a question may ask several other sub-questions, and the `oldcan:hasSubQuestion` object property captures this. Thus, the object property `oldcan:hasSubQuestion` relates one question with its sub-questions. Each question is linked to its associated answer. A question may have several answers collected from diverse sources. The `oldcan:hasAnswer` object property captures this relationship with its inverse `oldcan:isAnswerOf` property. Finally, a question is related

to a questionnaire with the `oldcan:isQuestionOf` object property where a single question is contained only in one questionnaire.

4.1.3. Answer

An answer is a written, spoken or illustrated response to a question. The answers to the questionnaires are collected using paper slips or forms. Here the domain experts are interested in modelling the information contained on the paper slips as answers. Earlier academic attempts sought to model lemmas as the only answers to the questions, due to a narrow focus on providing support to lexicographers who were interested in identifying headwords and a purely linguistic approach. This strategy was aimed at supporting lexicographers to extract the lemma from the answers and associate them with the paper slips and the questions in a separate paper slip record table. However, it ignores many of the collected answers other than the headwords, while several other questions have answers either in written, spoken or illustration formats. Depending on the type of question, the form of the answer varies, including sentences, individual words, multiword expressions, affixes, diagrams and drawings. For example, the answer to a thesaurus question is expected to be a word, or multiword expression, while the answer to an illustration question could be a sketch or a diagram.

In dboe@ema, the information about the answers is scattered in various tables including paper_slip_records, paper_slips and question tables. However, the paper_slip_records table is a significant table that links the questions, lemma and the paper slips, and it contains information which requires in-depth analysis before making a design decision. A closer look at the paper_slip_record table shows that it represents N-ary relations where a given record typically links a question, the corresponding paper slip and the extracted lemma (if any). Our recommended approach is to give an accurate semantics to this table and represent it as an Answer while keeping the name as an equivalent class in the ontology to support backward compatibility to the users of the system.

4.1.4. Paper Slips

A paper slip table represents both the medium and the information contained in individual paper slips. A paper slip may include various information related to the question including written or illustrated answers. The answer may vary depending on the type of question asked. A paper slip may further include the personal information related to the respondent, collector, place, and time. For the digitised paper slips, a scanned version of the data is also available as a media file.

4.1.5. Lemma, Multimedia, Source and Author

A lemma (`oldcan:lemma`) is a word that stands at the head of a definition in a dictionary. All the headwords in a dictionary are lemmas. A lemma in our ontology is represented as `oldcan:Lemma` and is linked to the Ontolex model. From a lexicographic point of view, a lemma is a complex entity which further includes several entities. Since modelling dictionary entries is well covered with existing ontologies, we reused such ontologies to represent the lemma semantically.

Multimedia (`oldcan:Multimedia`) refers to a medium that contains collected information. Any printed resource related to the entities have been digitised and stored in various formats as a multimedia file. There are various multimedia types contained in the database including Drawings, Audio, Video, Transparency, Photographs and Realia. Although we make a distinction between these media types, we do not discuss them here in detail. Questionnaires, Paper slips, Sources, Authors and Lemmas have corresponding multimedia files. `oldcan:Multimedia` is linked with the entities with the `oldcan:hasMultimedia/oldcan:isMultimediaOf` object properties.

A source (oldcan:Source) is anything that is used as a source of information which could be a person, a document or any other thing. Respondents back up their answers by citing their sources of information. We represented a source with many of its attributes linked to dc, Fabio [37] and other ontologies. A source is used by the respondents to support the answers and is linked to oldcan:paperslips with the oldcan:hasSource/oldcan:isSourceOf object properties.

We are further interested in the Authors of the questionnaires and, subsequently, the questions. Authors (oldcan:Author) in the collection could be individuals or organisational authors. Since the collection contains collectors, editors, and other persons who worked in the academy, we represent all these categories as FOAF:Agents while maintaining oldcan:Authors as a subclass of foaf:Agents. oldcan:Authors is linked to the questionnaire using the oldcan:hasAuthor/oldcan:isAuthorOf property. The resulting ontology in rdf/xml format can be downloaded from https://explorations4u.acdh.oeaw.ac.at/ontology/oldcan.

5. Semantic Uplift of Historical Resources at Exploreat

In recent years there has been a significant shift toward publishing data in a linked platform with the aim of serving users with a self-describing schema [38]. In our case, although the traditional collection is fully digitised using various formats, it is not yet available in an LOD format. The exploreAt! project has a broader objective of making the collection available for the public to support the exploration of the data. One of the approaches followed to support the exploration is to uplift the existing data into a suitable format together with the required semantics to interpret and use it independently.

Semantic uplift is a process of converting existing structured or semi-structured data into LOD based upon semantic-web technologies. These technologies heavily depend on existing standard vocabularies, domain-specific ontologies and W3C-recommended technologies such as RDF, RDFS, OWL, R2RML and SPARQL. The resulting LOD data is a graph database following the subject-predicate-object triple format that reuses existing vocabularies and ontologies to describe the target data with rich semantics. The semantic uplift process is a relatively new recommendation to publish data in the LOD platform; it leverages the generation of semantic data to support research.

We use semantic uplift in two phases. The first phase, which is presented in this paper, converts the data which is stored in a relational database (MySQL) to the RDF-format, whereas the second phase, which is a future work, converts the data which is still in TEI/XML format to an RDF-format [39–42]. Although the underlying principle in both phases is the same, the second phase requires additional analysis of the structure of the TEI/XML files to convert the data. The following section describes the first phase of semantic uplift in detail.

5.1. The Semantic Uplift Process

Similar to the semantic modelling process, this process requires a deep understanding of how the target data is structured and represented. The process involves data cleaning, mapping and LOD data generation (Figure 6).

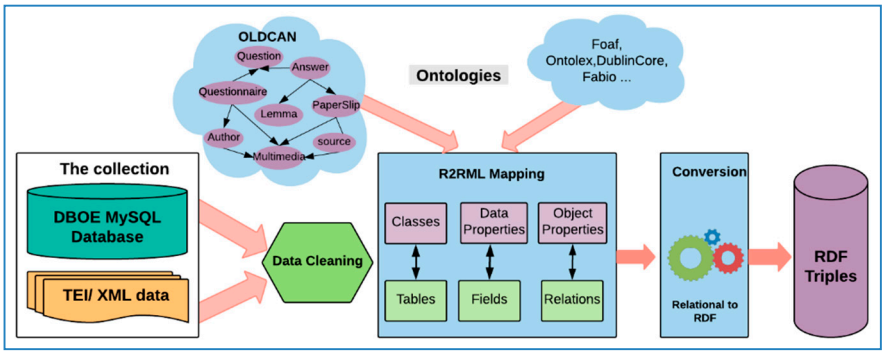

Figure 6. The workflow for the semantic uplift of DBÖ/dob@ema.

5.1.1. Data Cleaning

Data cleaning is one of the critical steps before publishing any data for public consumption. Data cleaning focuses on detecting and removing any inconsistencies and errors from the data [43]. The data at hand is not an exception in this regard as it contains missing values, invalid fields lacking

referential integrity etc. Another challenge is to distinguish between null values and empty values (sometimes tabs and whitespaces). Since the empty values are treated as values with empty data in the process of generating triples and are transformed into a meaningless statement (e.g., <http://exploreat.adaptcentre.ie/Questionnaire/1 oldcan:hasTitle " ">), all the empty values need to be converted into a proper null value. To achieve this, we run a batch script that converts the empty values to null values in MySQL. Then, we identified fields that contain null values across all the records and removed them from the mapping. The data also contains invalid values that are introduced either during the data conversion stage or during data processing stages. By consulting the domain experts, we cleaned the records and restored the fields to the original data. For example, all questionnaire topics contain the word "Fragebogen. X" before the title of the questionnaires. This field is cleaned by removing "Fragebogen. X". from the title field where X stands for the questionnaire number. Some missing data is also repopulated from other internal records. A good example is the inclusion of the authors of the questionnaires. Although we cleaned most of the technical and syntactic errors in the data, we did not manage to maintain the robustness of the data, and cleaning the semantic errors became very difficult and time-consuming.

One of the potential approaches to reduce semantic errors is to actively engage citizen scientists in reporting back whenever they encounter such errors. Identifying the semantic errors requires a fair level of knowledge and willingness from citizen scientists; however, it has a great potential to bring such errors to our attention. Another approach is to use Shape Expression (ShEx) language [44] to validate the conformance of the generated data against certain constraints. This will enable us to catch some of the surface-level semantic errors. However, this requires identifying such semantic errors and representing them using rules in ShEx language. Finally, the use of machine learning to classify individual instances of the data could contribute towards ensuring the quality of the data where outliers will be identified and evaluated manually. The machine learning approach involves generating training data containing both negative and positive examples which could be used to train the system. In a supervised environment, the experts could provide training datasets which will require a good amount of the expert's time. This challenge works as a case study introducing collaborative lexicography and crowd innovation.

5.1.2. R2RML Mapping

There are various methods and tools used to transform relational databases to a semantically compatible format including direct mapping [45] and domain semantics-driven mapping [46]. We followed R2RML [47] 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 external domain ontology in R2RML. Since R2RML is a vocabulary by itself, it stores the mappings from a relational database as RDF files and allows the inclusion of provenance information which facilitates knowledge discovery and reuse of mappings. In addition to mapping from a relational database to RDF, R2RML serves the purpose of mapping back from RDF to a relational database which makes it suitable for reverse engineering purposes. However, it requires more effort compared to direct mapping. R2RML is used to map the relational data into an LOD. This phase includes the following steps:

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

The mapping of the questionnaire and question entities and their fields is presented in Table 5. A questionnaire table is transformed into a view using an SQL statement that decodes the questionnaire type from the id and assigns the type to the respective questionnaires. It uses the URL http://exploreat.adaptcentre.ie/Questionnaire/\protect\T1\textbraceleftid\protect\

T1\textbraceright to generate a resource for each individual questionnaire. This means that a row in the table (e.g., questionnaire 1) is identified by a fixed URL http://exploreat.adaptcentre.ie/Questionnaire/1 and its type is assigned as both oldcan:questionnaire and oldcan:SystematicQuestionnaire. Then the questionnaire will also have other triples that describe the information in the columns. The mapping further generates the links between the questionnaire and the author of a questionnaire using the oldcan: has Author property (Table 6).

Table 5. A sample R2RML mapping of questionnaire and question.

Questionnaire	Question
<#QuestionnaireTriplesMap> a rr:TriplesMap; rr:logicalTable [rr:sqlQuery """ SELECT Fragebogen.*, (CASE fragebogen_typ_id WHEN '1' THEN 'SystematicQuestionnaire' WHEN '2' THEN 'AdditionalQuestionnaire' WHEN '3' THEN 'DialectographicQuestionnaire' END) QUESTIONNAIRETYPE FROM Fragebogen """]; rr:subjectMap [rr:template "http://exploreat.adaptcentre.ie/Questionnaire/\protect\T1\ textbraceleftid\protect\T1\textbraceright" ; rr:class oldcan:Questionnaire ; rr:graph <http://exploreat.adaptcentre.ie/Questionnaire_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:title ; rr:predicate rdfs:label; rr:objectMap [rr:column "titel" ; rr:language "de";] ; rr:graph <http://exploreat.adaptcentre.ie/Questionnaire_graph> ;] ; rr:predicateObjectMap [rr:predicate oldcan:publicationYear ; rr:objectMap [rr:column "erscheinungsjahr"] ; rr:graph <http://exploreat.adaptcentre.ie/Questionnaire_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:note ; rr:objectMap [rr:column "anmerkung"; rr:language "de"]; rr:graph <http://exploreat.adaptcentre.ie/Questionnaire_graph> ;]; rr:predicateObjectMap [rr:predicate rdf:type ; rr:objectMap [rr:template "https://explorations4u.acdh.oeaw.ac.at/ontology/oldcan#{QUESTIONNAIRETYPE}"; rr:graph <http://exploreat.adaptcentre.ie/Questionnaire_graph> ;] ;]; rr:predicateObjectMap [rr:predicate oldcan:hasAuthor ; rr:predicate dct:creator; rr:graph <http://exploreat.adaptcentre.ie/Questionnaire_graph> ; rr:objectMap [rr:parentTriplesMap <#PersonTripleMap> ; rr:joinCondition [rr:child "person_id" ; rr:parent "id" ;]]] ;.	<#QuestionTripleMap> a rr:TriplesMap; rr:logicalTable [rr:sqlQuery """SELECT Frage.* FROM Frage"""]; rr:subjectMap [rr:template "http://exploreat.adaptcentre.ie/Question/ \protect\T1\textbraceleftid\protect\T1\textbraceright" ; rr:class oldcan:Question ; rr:graph <http://exploreat.adaptcentre.ie/Question_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:combinedID ; rr:objectMap [rr:template"http://exploreat.adaptcentre.ie/Question/ \protect\T1\textbraceleftfragebogen_id\protect\T1\ textbraceright-\protect\T1\textbraceleftnummer\protect\ T1\textbraceright"]; rr:graph <http://exploreat.adaptcentre.ie/Question_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:number ; rr:objectMap [rr:column "nummer"] ; rr:graph <http://exploreat.adaptcentre.ie/Question_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:originalQuestion ; rr:predicate rdfs:label; rr:objectMap [rr:column "originalfrage" ; rr:language "de";] ; rr:graph <http://exploreat.adaptcentre.ie/Question_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:shortQuestion ; rr:objectMap [rr:column "kurzfrage" ; rr:language "de";] ; rr:graph <http://exploreat.adaptcentre.ie/Question_graph> ;]; rr:predicateObjectMap [rr:predicate oldcan:isQuestionOf ; rr:graph <http://exploreat.adaptcentre.ie/Question_graph> ; rr:objectMap [rr:parentTriplesMap <#QuestionnaireTriplesMap> ; rr:joinCondition [rr:child "fragebogen_id" ; rr:parent "id" ;]]] ;.

Table 6. Sample triples for questionnaire 1.

Generated Triples for Questionnaire 1
(http://exploreat.adaptcentre.ie/Questionnaire/1 rdfs:type oldcan:Questionnaire)
(http://exploreat.adaptcentre.ie/Questionnaire/1 rdfs:type oldcan:SystematicQuestionnaire)
(http://exploreat.adaptcentre.ie/Questionnaire/1 oldcan:publicationYear 1920)
(http://exploreat.adaptcentre.ie/Questionnaire/1 oldcan:title "Kopf")
(http://exploreat.adaptcentre.ie/Questionnaire/1 oldcan:hasAuthor "http://exploreat.adaptcentre.ie/Person/22192"

5.2. Linked Open Data (LOD) Generation

The tables in the relational database are converted into classes, and the fields are converted into triples of those classes. We used r2rml [48] to generate the LOD data (exploreAT! Dublin City University, Adapt Centre: http://exploreat.adaptcentre.ie/#APIs) automatically. A snapshot of the triples is presented in Table 7.

Table 7. A snapshot of the RDF triples generated from the data.

Subject	Predicate	Object
http://exploreat.adaptcentre.ie/Questionnaire/1	rdfs:label	"Kopf (1)"@de
http://exploreat.adaptcentre.ie/Questionnaire/1	oldcan:note	"resfb1"@de
http://exploreat.adaptcentre.ie/Questionnaire/1	rdf:type	oldcan:Questionnaire
http://exploreat.adaptcentre.ie/Questionnaire/1	rdf:type	oldcan:SystematicQuestionnaire
http://exploreat.adaptcentre.ie/Questionnaire/1	oldcan:title	"Kopf (1)"@de
http://exploreat.adaptcentre.ie/Questionnaire/1	oldcan:hasAuthor	http://exploreat.adaptcentre.ie/Person/22192
http://exploreat.adaptcentre.ie/Questionnaire/1	oldcan:publicationYear	1920
http://exploreat.adaptcentre.ie/Question/1	rdfslabel	"Kopf: Kopf, Haupt; auch scherzh./übertr."@de
http://exploreat.adaptcentre.ie/Question/1	rdf:type	oldcan:Question
http://exploreat.adaptcentre.ie/Question/1	oldcan:isQuestionOf	http://exploreat.adaptcentre.ie/Questionnaire/1
http://exploreat.adaptcentre.ie/Question/1	oldcan:combinedID	http://exploreat.adaptcentre.ie/Question/1-A1
http://exploreat.adaptcentre.ie/Question/1	oldcan:number	"A1"
http://exploreat.adaptcentre.ie/Question/1	oldcan:originalQuestion	"Kopf: Kopf, Haupt; auch scherzh./übertr."@de
http://exploreat.adaptcentre.ie/Question/1	oldcan:shortQuestion	"Kopf, Haupt; auch scherzh./übertr."@de

6. Implementation and Validation

To support the immediate requirements of the domain experts, we focus on uplifting the core entities in the collection. These core entities cover the most significant part of the user's information requirements, and thus are used to evaluate the quality of the resulting data using exploration paths frequently used by the users. The resulting data is organised into eight named graphs which will enable us to answer queries efficiently. A summary of the number of triples generated in each of the named graphs is presented in Table 8.

Table 8. Distribution of triples in across the named graphs.

Named Graph	Unique Entities	Triples	Named Graph	Unique Entities	Triples
Questionnaire_graph	762	2969	Source_graph	16839	231537
Question_graph	24382	163705	Agents_graph	11163	123438
Paperslip_graph	65839	539749	Multimedia_graph	8218	63741
Papersliprecord_graph	140509	824925	Lemma_graph	61878	921261

6.1. Exploration Paths

The domain experts who are working on supporting the requirements of lexicographers, linguists, historians and citizen scientists collected several types of information requirements. These requirements are summarized, and representative questions are identified to evaluate the resulting semantic data qualitatively (Table 9). We evaluate the ontology in terms of assisting the analysis of the queries and representing them in a structured manner using a SPARQL query, and the accuracy of the resulting data in terms of identifying the required data.

Table 9. Queries extracted from exploration paths.

Query	Description	Purpose
Q1	All the questionnaires that deal with a topic T	Conceptualisation and topic discovery
Q2	All the questionnaires whose author has a gender G (male, female, unknown)	Biographical and prosopographical analysis
Q3	All the paper slips that contain answers to question X	Generic, historical and cultural
Q4	Number of questions authored by a female author	Statistical analysis
Q5	All the questions that are related to a lemma x	Lexical and lexicographic analysis
Q6	All answers that are collected for questionnaire X	Generic, historical and cultural inquiry

These queries are translated into SPARQL queries using the inputs from the ontology and the mapping. The ontology is used to link the entities in the queries, and the mapping is used to construct the SPARQL graph pattern. The resulting SPARQL queries are presented in Table 10. The queries are qualitatively evaluated for their accuracy by both users and experts by comparing them with the results gained from MySQL.

To further describe the validation process and the usability of the system, we now discuss in detail some of the queries outlined above. The queries are representative of the widely asked questions by lexicographers for various reasons. One of the rationales behind such questions is the need to understand the usage of specific terms, including the context in which the terms are used and the various forms they have in different contexts. One way of answering this question is to search the collection using questions that are related to that specific term and what aspects of the term are collected via the paper slips. This information allows users to explore not only the questions, but also the timeline of the data collection, and the evolution of the terms at various stages as paper slips containing answers to that specific question vary in the time and place of collection. Query 5, for example, answers such a question and provides users with the questionnaires and questions that contain the terms, and links the questions with their answers as well as collectors and several pieces of additional information that could be interesting for the users. For example, the query that searches for colour ("Farbe" in German) resulted in all the questionnaires that deal with colours and provides links to analyse several aspects of colour such as making colours, colour composition, food colour, colour in festivals and jokes. This query resulted in two major questionnaires containing several questions related to colour. Questionnaire 12 (<http://exploreat.adaptcentre.ie/Questionaire/12>, "Nase, ohr") which is not directly related to colour but has an answer which mentions the word colour. Questionnaire 4 (<http://exploreat.adaptcentre.ie/Questionaire/4>, "Kopf") also contains one answer related to colour but focusing on other aspects of colour which is related to a body part (head). Both results open an exploration path to individual questions in the questionnaires. Since we provide support of a browsable API, users can navigate to each of the individual questions contained in the questionnaires and see detailed questions where colour is mentioned as part of the lemma related to the answer. Furthermore, they can navigate to the answers associated with these questions.

While Query 5 begins exploration from the lemma and tries to find all relevant questions whose answer includes the word colour, Query 1 provides a top-down exploration which results in the identification of different sets of questionnaires that have colour as their primary topic. Using "Farbe" as a parameter in Query 1, we primarily obtain Questionnaire 53 which deals with colours (<http://exploreat.adaptcentre.ie/Questionaire/53>, "Farbe") and Questionnaire 393 which deals with "flowers, colours and trees" (<http://exploreat.adaptcentre.ie/Questionaire/393>, "Blumen, Farben und Bäume"). Providing such flexibility in searching the collections from a different angle enabled the users to explore the data effectively and allowed them to pose various queries related to the collection.

Domain experts evaluated the results of these queries and compared the answers with the answers gained from the legacy systems. The results show that these queries were able to retrieve the same result but with much more semantics associated. The results are self-explanatory in that there is always an associated ontology which provides a standard description of the entities and the relationships between the entities.

The domain experts need to consult various tables before answering these questions and, furthermore, need to explain the results in detail. Using SPARQL, it is possible that these questions could be answered directly employing the SPARQL endpoint and through the provided API. Using the API, users could easily navigate through different entities following the available links. Although MySQL supports structured queries, browsing the entities using the semantic links is achieved through a browsable API that consumes the LOD. The API enables non-technical users to browse the data starting from any of their chosen entities and click on any of the links to move to linked entities. It also supports autonomous systems to access any required data via an http request. This is demonstrated by the integration of the API with a visualisation tool that allows us to navigate

through the links to discover further interesting links [49]. Finally, the resulting LOD is linked to existing resources such as DBpedia. The model at this stage supports a bilingual description of the entities and attributes.

Table 10. SPARQL queries for navigation paths.

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> prefix oldcan: <https://explorations4u.acdh.oeaw.ac.at/ontology/oldcan#>	prefix xsd: <http://www.w3.org/2001/XMLSchema#> PREFIX foaf: <http://xmlns.com/foaf/0.1/>

Q1
SELECT * FROM
Named <http://exploreat.adaptcentre.ie/Questionnaire_graph>
WHERE {
Graph <http://exploreat.adaptcentre.ie/Questionnaire_graph>
{
?subject rdfs:label ?object.
FILTER regex(?object,"Religion und Kirch", "i")
} }
	Q2
	SELECT ?questionnaire
	FROM
	<http://exploreat.adaptcentre.ie/Questionnaire_graph>
	FROM <http://exploreat.adaptcentre.ie/Person_graph>
	WHERE {
	?questionnaire oldcan:hasAuthor ?author.
	?author foaf:gender "Female".
	}

Q3
SELECT ?paperSlip
FROM <http://exploreat.adaptcentre.ie/Question_graph>
FROM <http://exploreat.adaptcentre.ie/PaperSlipRecord_graph>
FROM <http://exploreat.adaptcentre.ie/PaperSlip_graph>
WHERE {
?paperSlipRecord oldcan:containsQuestion
<http://exploreat.adaptcentre.ie/Question/53>.
?paperSlipRecord oldcan:hasPaperSlip ?paperSlip.
}
	Q4
	SELECT count(distinct ?question)
	FROM
	<http://exploreat.adaptcentre.ie/Questionnaire_graph>
	FROM <http://exploreat.adaptcentre.ie/Question_graph>
	FROM <http://exploreat.adaptcentre.ie/Person_graph>
	WHERE {
	?questionnaire oldcan:hasAuthor ?author.
	?author foaf:gender "Female".
	?question oldcan:isQuestionOf ?questionnaire.
	}

Q5
SELECT distinct ?questionnaire ?text
FROM <http://exploreat.adaptcentre.ie/Questionnaire_graph>
FROM <http://exploreat.adaptcentre.ie/Question_graph>
FROM <http://exploreat.adaptcentre.ie/Lemma_graph>
FROM <http://exploreat.adaptcentre.ie/PaperSlipRecord_graph>
WHERE {
?question oldcan:isQuestionOf ?questionnaire.
?paperSlipRecord oldcan:containsQuestion ?question.
?paperSlipRecord oldcan:hasLemma ?lemma.
?lemma rdfs:label ?text.
FILTER regex(?text, "Lempe", "i")
}
	Q6
	SELECT distinct ?papersliprecord
	FROM <http://exploreat.adaptcentre.ie/Question_graph>
	FROM
	<http://exploreat.adaptcentre.ie/PaperSlipRecord_graph>
	FROM <http://exploreat.adaptcentre.ie/Lemma_graph>
	WHERE {
	?question oldcan:isQuestionOf
	<http://exploreat.adaptcentre.ie/Questionnaire/1>.
	?papersliprecord oldcan:containsQuestion ?question.
	?papersliprecord oldcan:hasLemma ?lemma.
	}

6.2. Qualitative Evaluation of the Ontology Model

Research indicates that ontology models are widely evaluated for their accuracy, completeness, adaptability, consistency and other additional factors [50]. Experts in the project evaluated the resulting ontology model for its accuracy and completeness. The accuracy measures whether the identified concepts, properties and axioms comply with the domain knowledge. In this validation process, domain experts were presented with each candidate entity together with their corresponding properties and definitions using a spreadsheet in a collaborative editing environment. The domain experts and ontology engineers had several discussions and reached a consensus after evaluating each entity and cross-examining all the supporting evidence. Supporting evidence was drawn both from the

instances in the database and additional published and unpublished resources held in the academy. Completeness focuses on the model's coverage of the domain as outlined in the initial requirement. Domain experts agree that the current version of the model covers most of the core entities of major interest in the collection. One shortcoming in the coverage is that more entities that were not identified in the initial requirement became visible and proved essential to describe the core entities. These entities require in-depth analysis and further modelling, which compromised the complete coverage of the domain. In general, the experts agree to limit the focus of the model only to the initial requirement. Thus, even if the current version of the model is complete concerning the initial requirement, future improvements are unavoidable.

Apart from the evaluation undertaken by the domain experts, the ontology is evaluated for its consistency to make sure that it is free from any contradiction. The resulting ontology is evaluated for its consistency using Fact ++1.6.5 [51] and Hermit 1.3.8.413 [52] reasoners, which are bundled with the Protege ontology editor, where we found no contradiction in the ontology. We further evaluated our model using an online tool (OOPS) which evaluates ontologies using several evaluation criteria. The evaluation of the model for its adaptability to other similar domains has not yet been conducted but is considered as one of the tasks for future work.

6.3. Qualitative Evaluation of the Mapping

Further to the evaluation of the model, the mapping of the data using R2RML was evaluated for its accuracy and completeness. The mapping transformed the data in MySQL tables to a triple by attaching rich semantics from the ontology model. The mapping is validated by comparing the descriptions of the tables and the columns with the description of the properties in the ontology. Some of the attributes in the database tables are enriched using two or more properties that make the resulting data more interlinked.

The final system is validated against the user requirements identified in Section 3.1. Domain experts assess the resulting model as to whether it is capable of accurately representing the entities, attributes and relationships. Although all the available attributes are not deemed to be significant due to the incompleteness of the data, those that are identified by domain experts are included and represented in the model. These entities are used to annotate the data in the MySQL database and are used in the semantic enrichment process. One drawback is that, in some cases, the data contained in a specific table is not homogeneous, which as a result proved to be a challenge to map the table directly to the corresponding entity. This requires (i) an in-depth analysis of tables that contain heterogeneous data, and (ii) that the model be evolved to fit the requirements.

7. Related Work

Even if the breadth and the depth of analysis vary, there has been similar research conducted elsewhere in the area of analysis of culture using linguistic constructs. The authors in [53] perform a quantitative analysis of culture using text extracted from millions of digitised books. The focus of their research is extensive in that they collected more than 5 million digitised books and analysed the linguistic and cultural phenomena observed in the English language. They performed a quantitative analysis and observed the distribution of words over a long period based on significant events in history. An exciting aspect of their research which aligns with our objective is the analysis of words and their evolution over time. Even if we have a similar objective to achieve by the end of the project, in our work to date, our primary focus has been to enrich the collection semantically and make it available for similar analysis in the future. We are also interested in qualitative analysis to provide supporting evidence for cultural linguists, lexicographers and citizen scientists. Our dataset also differs in that we deal with raw data collected directly from respondents that did not pass through extensive editing, unlike the content of published books.

Another closely related work is conducted by Strok et al. [21] focusing on a semantic annotation of a natural history collection. This research dealt with a large collection of historical biodiversity

expeditions housed in several European natural history archives. The authors understood the need for a semantic model to annotate such domains and analysed various existing models before proposing a new model. They focused on 8000 field book pages and annotated the content of the pages using the proposed semantic model. The approach we followed in our work is similar to the one they followed to build the semantic model, in that the method relies significantly on analysis of the primary resources and the knowledge of the domain experts. Our approach differs from theirs in its focus. The focus of our work is not extracting named entities but enriching the entities in the collection and linking the individual instances with rich semantics. While they use their model to annotate the contents of the pages of the books semantically, we use our model to enrich semantically the records that are already in the relational database.

Annotation of cultural heritage collections has been done from various perspectives. Guus, et al. [54] described their semantic web application for semantic annotation and search of large cultural-heritage objects from various public museums. They combined existing vocabularies to describe the collection but do not produce new semantic models. The resulting semantic annotation was used to support semantic search and to organise search results using predefined categories. Unlike their semantic annotation work which benefits from existing generic ontologies, we believe that the use of domain-specific ontologies in addition to generic semantic models provides a deeper understanding of the collection beyond describing the metadata of the content. Our work focuses on building a semantic model to describe the data collection methods in order to fill the existing gap in the modelling of questionnaires, questions, answers and related entities used in sociocultural linguistics. Their work focused on metadata level annotation of cultural objects, whereas we focus on domain-specific annotation of both the collected resources and the data collection methods.

With a similar objective of making historical resources accessible, authors in [20] employed semantic enrichment of a multilingual cultural heritage archive with an LOD. Their case study on the Historische Kranten involved digitisation and OCR processing and publication of millions of articles published between 1818 and 1927. To support semantic enrichment, knowledge extraction and entity linking are used, and named entities that are identified in the collection are linked with DBpedia. The approach used in the mapping mainly focuses on linking the named entities with the DBpedia URIs. This approach is a popular method which is used to disambiguate and enrich documents semantically even when tools supporting the process are available [35].

Finally, initiatives to convert existing lexicographic resources to a linguistic LOD are proposed [55,56]. The use of generic models for semantic annotation continues to dominate the area, while the use of domain-specific ontologies provides a fine-grained semantics but with a higher amount of budget and expertise. This work tries to achieve semantic annotation at a fine-grained level with the involvement of domain experts.

8. Discussion and Conclusions

The effort to open up legacy collections to make them accessible, usable and searchable has increased with the development of LOD platforms that facilitate the publication of a wide range of content. For domain-specific content, developing semantic models that describe the domain of interest is crucial. This paper presented, first, a semantic model for enriching and publishing traditional data of Bavarian dialects. We argue that the development of the model facilitates the semantic publication of the data on the LOD platform and facilitates the exploitation of the data from various points of views. It further paves the way for researchers to understand and compare the conceptualisation of entities from different perspectives as well as their evolution over time. Second, we presented the mapping of the data held in a relational database to an LOD format using a domain-specific semantic model. The mapping enables the semantic enrichment of the collection and the interlinking among entities within the collection and with external resources such as DBpedia. Third, we presented the resulting LOD which is made available via a downloadable file, a browsable API and a SPARQL endpoint.

The qualitative evaluation demonstrated that the developed ontology (OLDCAN) complies with the requirements of the domain experts and users. We demonstrated that the resulting ontology is usable to semantically enrich the data. It provides rich semantics, interoperability, structure and facility to interlink the collection. However, the model will extend to include additional entities as we cover the additional data which is in the TEI/XML format. The exploration paths demonstrated that LOD is also usable by domain experts for answering their frequently asked queries by building exploration paths and interlinking the data with external resources. The resulting LOD further supports browsing and visual exploration of the data, which provides additional methods of interacting with the collection.

One of the challenges of the ontology development and the semantic uplifting process is that, while the process opens a new and better way of exploring the data, we also discovered new requirements that were not captured initially. This is both a challenge and an opportunity in that the more insight we gain from the data, the better we model the system, but it is always tricky to decide when to stop adding new information in the model.

Our future work will focus on two key areas. The first one is to convert the wealth of information contained in the TEI/XML files, which involves the identification of additional entities, semantics and the inclusion of a large dataset. The second direction will focus on providing various ways of exploring the data in addition to the browsable API and the SPARQL queries. We aim to build visualisation tools to support user requirements via an automated composition of exploration paths. Since our users have diverse interests in the data, supporting their requirements with visual analytics is crucial. Finally, we will focus on introducing a feedback loop for our users to contribute new knowledge to the system.

Finally, our effort to enrich the collection with state-of-the-art semantics and make it available in various formats for both expert users, citizen scientists and autonomous machines will facilitate efficient exploration and exploitation of traditional data of historical and linguistic importance.

Author Contributions: Conceptualization, Y.A. and E.W.-V.; Data curation, Y.A., A.D. and B.P.; Formal analysis, Y.A.; Investigation, Y.A., A.D. and B.P.; Methodology, Y.A.; Project administration, Y.A., A.D., E.W.-V. and A.W.; Resources, Y.A. and B.P.; Software, Y.A.; Supervision, E.W.-V. and A.W.; Validation, Y.A., A.D. and B.P.; Visualization, Y.A.; Writing—original draft, Y.A., A.D. and B.P.; Writing—review and editing, Y.A., E.W.-V. and A.W.

Funding: 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.

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

References

1. Doerr, M. Ontologies for Cultural Heritage. In *Handbook on Ontologies*; Springer: Berlin/Heidelberg, Germany, 2009.
2. Kansa, E.C.; Kansa, S.W.; Burton, M.M.; Stankowski, C. Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies* **2010**, *6*, 301–326. [CrossRef]
3. Beretta, F.; Ferhod, D.; Gedzelman, S.; Vernus, P. The SyMoGIH project: Publishing and sharing historical data on the semantic web. In Proceedings of the Digital Humanities 2014, Lausanne, Switzerland, 7–12 July 2014; pp. 469–470.
4. Meroño-Peñuela, A.; Ashkpour, A.; Van Erp, M.; Mandemakers, K.; Breure, L.; Scharnhorst, A.; Schlobach, S.; Van Harmelen, F. Semantic Technologies for Historical Research: A Survey. *Semant. Web* **2015**, *6*, 539–564. [CrossRef]
5. Lampron, P.; Mixter, J.; Han, M.J.K. Challenges of mapping digital collections metadata to Schema.org: Working with CONTENTdm. In Proceedings of the 10th International Research Conference on Metadata and Semantics Research, Göttingen, Germany, 22–25 November 2016; pp. 181–186.

6. Wandl-Vogt, E.; Kieslinger, B.; O´Connor, A.; Theron, R. exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. Available online: http://docplayer.org/16597238-Exploreat-perspektiven-einer-transformation-am-beispiel-eines-lexikographischen-jahrhundertprojekts.html (accessed on 22 November 2018).
7. Wandl-Vogt, E. Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema). Available online: https://wboe.oeaw.ac.at/projekt/beschreibung/ (accessed on 22 November 2018).
8. Dominque, J.; Fensel, D.; Hendler, J.A. *Handbook of Semantic Web Technologies*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
9. Nevalainen, T.; Raumolin-Brunberg, H. Historical Sociolinguistics: Origins, Motivations, and Paradigms. In *The Handbook of Historical Sociolinguistics*; Wiley-Blackwell: Hoboken, NJ, USA, 2012; pp. 22–40.
10. Kramsch, C.; Widdowson, H. *Language and Culture*; Oxford University Press: Oxford, UK, 1998.
11. Chiarcos, C.; Cimiano, P.; Declerck, T.; McCrae, J.P. Linguistic Linked Open Data (LLOD)—Introduction and Overview. In Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data, Pisa, Italy, 23 September 2013.
12. Burnard, L. *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*; OpenEdition Press: Marseille, France, 2014.
13. McCrae, J.P.; Bosque-Gil, J.; Gracia, J.; Buitelaar, P.; Cimiano, P. The OntoLex-Lemon Model: Development and applications. In Proceedings of the the 5th Biennial Conference on Electronic Lexicography (eLex 2017), Leiden, The Netherlands, 19–21 September 2017; pp. 587–597.
14. Pedersen, B.; McCrae, J.; Tiberius, C.; Krek, S. ELEXIS—A European infrastructure fostering cooperation and information exchange among lexicographical research communities. In Proceedings of the 9th Global WordNet Conference, Singapore, 8–12 January 2018.
15. Wandl-Vogt, E. Wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I Dinamlex) (mit 10 Abbildungen). In *Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen*; Praesens: Vienna, Austria, 2008; pp. 93–112.
16. Arbeitsplan. Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912. Karton 1. In *Arbeitsplan-a-h Bayerisch-Österreichisches Wörterbuch*; Archive of the Austrian Academy of Sciences: Vienna, Austria, 1912.
17. WBÖ. Wörterbuch der bairischen Mundarten in Österreich (1970–2015). In *Bayerisches Wörterbuch: I. Österreich, 5 vols. Ed.*; Verlag der Österreichischen Akademie der Wissenschaften: Vienna, Austria.
18. Barabas, B.; Hareter-Kroiss, C.; Hofstetter, B.; Mayer, L.; Piringer, B.; Schwaiger, S. Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In *Fokus Dialekt. Analysieren-Dokumentieren-Kommunizieren*; Olms Verlag: Hildesheim, Germany, 2010; pp. 47–64.
19. Schopper, D.; Bowers, J.; Wandl-Vogt, E. dboe@TEI: Remodelling a data-base of dialects into a rich LOD resource. In Proceedings of the 9th International Conference on Tangible, Embedded, and Embodied Interaction (TEI 2015), Stanford, CA, USA, 15–19 January 2015.
20. De Wilde, M.; Hengchen, S. Semantic Enrichment of a Multilingual Archive with Linked Open Data. *Digit. Hum. Q.* **2017**, *11*, 1938–4122.
21. Strok, L.; Weber, A.; Miracle, G.G.; Verbeek, F.; Plaat, A.; Herik, J.V.D.; Wolstencroft, K. Semantic annotation of natural history collections. *Web Semant. Sci. Serv. Agents World Wide Web* **2018**, in press. [CrossRef]
22. Hrastnig, E. A Linked Data Approach for Digital Humanities. Prototypical Storage of a Dialect Data Set in a Triplestore. Master's Thesis, Graz University of Technology, Graz, Austria, January 2017.
23. Peroni, S. Automating Semantic Publishing. *Data Sci.* **2017**, *1*, 155–173. [CrossRef]
24. Scholz, J.; Lampoltshammer, T.J.; Bartelme, N.; Wandl-Vogt, E. Spatial-temporal Modeling of Linguistic Regions and Processes with Combined Indeterminate and Crisp Boundaries. In *Progress in Cartography: EuroCarto 2015*; Gartner, G., Jobst, M., Huang, H., Eds.; Springer: Cham, Switzerland, 2016; pp. 133–151.
25. Scholz, J.; Hrastnig, E.; Wandl-Vogt, E. A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In Proceedings of the Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017), L'Aquila, Italy, 4–8 September 2017; pp. 275–282.
26. Boyce, S.; Pahl, C. Developing Domain Ontologies for Course Content. *Educ. Technol. Soc.* **2007**, *10*, 275–288.

27. Noy, N.F.; Mcguinness, D.L. Ontology Development 101: A Guide to Creating Your First Ontology. Available online: http://www.corais.org/sites/default/files/ontology_development_101_aguide_to_creating_your_first_ontology.pdf (accessed on 22 November 2018).
28. Gura, C.; Piringer, B.; Wandl-Vogt, E. Nation Building durch Großlandschaftswörterbücher. Das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) als identitätsstiftender Faktor des österreichischen Bewusstseins. Status (unpublished).
29. Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data—The Story So Far. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–22. [CrossRef]
30. Hogan, A.; Umbrich, J.; Harth, A.; Cyganiak, R.; Polleres, A.; Decker, S. An empirical survey of linked data conformance. *Web Semant. Sci. Serv. Agents World Wide Web* **2012**, *14*, 14–44. [CrossRef]
31. Uschold, M.; Gruninger, M. Ontologies: Principles, methods, and applications. *Knowl. Eng. Rev.* **1996**, *11*, 93–155. [CrossRef]
32. Edgar, S.M.; Alexei, S.A. Ontology for knowledge management in software maintenance. *Int. J. Inf. Manag.* **2014**, *34*, 704–710.
33. Brickley, D.; Miller, L. FOAF Vocabulary Specification 0.99, 2014. Namespace Document. Available online: http://xmlns.com/foaf/spec/ (accessed on 23 November 2018).
34. Board, DCMI Usage. DCMI Metadata Terms, 2014. Dublin Core Metadata Initiative. Available online: http://dublincore.org/documents/dcmi-terms/ (accessed on 23 November 2018).
35. Mendes, P.N.; Jakob, M.; García-Silva, A.; Bizer, C. DBpedia spotlight. In Proceedings of the 7th International Conference on Semantic Systems, Graz, Austria, 7–9 September 2011; pp. 1–8.
36. Abgaz, Y.; Dorn, A.; Piringer, B.; Wandl-Vogt, E.; Way, A. A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 21–29.
37. Shotton, D.; Peroni, S. FaBiO, the FRBR-aligned Bibliographic Ontology, 2018. Available online: https://sparontologies.github.io/fabio/current/fabio.html (accessed on 23 November 2018).
38. Heath, T.; Bizer, C. Semantic Annotation and Retrieval: Web of Data. In *Handbook of Semantic Web Technologies*; Domingue, J., Fensel, D., Hendler, J.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2011.
39. Ferdinand, M.; Christian, Z.; David, T. Lifting XML Schema to OWL. In Proceedings of the Web Engineering—4th International Conference (ICWE 2004), Munich, Germany, 26–30 July 2004; pp. 354–358.
40. Battle, S. Gloze: XML to RDF and back again. In Proceedings of the First Jena User Conference, Bristol, UK, 10–11 May 2006.
41. Deursen, D.V.; Poppe, C.; Martens, G.; Mannens, E.; Walle, R.V.d. XML to RDF Conversion: A Generic Approach. In Proceedings of the 2008 International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution, Florence, Italy, 17–19 November 2008; pp. 138–144.
42. Simpson, J.; Brown, S. From XML to RDF in the Orlando Project. In Proceedings of the International Conference on Culture and Computing. Culture and Computing, Kyoto, Japan, 16-18 September 2013; pp. 194–195.
43. Gueta, T.; Carmel, Y. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecol. Inform.* **2016**, *34*, 139–145. [CrossRef]
44. Prud'hommeaux, E.; Labra Gayo, J.E.; Solbrig, H. Shape Expressions: An RDF Validation and Transformation Language. In Proceedings of the 10th International Conference on Semantic Systems (Sem2014), Leipzig, Germany, 14 July 2014; pp. 32–40. [CrossRef]
45. Berners-Lee, T. Relational Databases on the Semantic Web. In Design Issues for the World Wide Web. Available online: https://www.w3.org/DesignIssues/RDB-RDF.html (accessed on 22 November 2018).
46. Michel, F.; Montagnat, J.; Faron, Z.C. A Survey of RDB to RDF Translation Approaches and Tools. Available online: https://hal.archives-ouvertes.fr/hal-00903568v1 (accessed on 22 November 2018).
47. Das, S.; Sundara, S.; Cyganiak, R. R2RML: RDB to RDF Mapping Language. W3C RDB2RDF Working Group. Available online: https://www.w3.org/TR/r2rml/ (accessed on 23 November 2018).
48. Debruyne, C.; O'Sullivan, D. R2RML-F: Towards Sharing and Executing Domain Logic in R2RML Mappings. In Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with the 25th International World Wide Web Conference (WWW 2016), Montreal, QC, Canada, 12 April 2016.

49. Dorn, A.; Wandl-Vogt, E.; Abgaz, Y.; Benito Santos, A.; Therón, R. Unlocking Cultural Knowledge in Indigenous Language Resources: Collaborative Computing Methodologies. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
50. Raad, J.; Cruz, C. A Survey on Ontology Evaluation Methods. In Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 12–14 November 2015; pp. 179–186.
51. Tsarkov, D.; Horrocks, I. FaCT Description Logic Reasoner: System Description. In Proceedings of the International Joint Conference on Automated Reasoning, Seattle, WA, USA, 17–20 August 2006; pp. 292–297.
52. Glimm, B.; Horrocks, I.; Motik, B.; Stoilos, G.; Wang, Z. HermiT: An OWL 2 Reasoner. *J. Autom. Reason.* **2014**, *53*, 245–269. [CrossRef]
53. Jean-Baptiste, M.; Yuan, K.S.; Aviva, P.A.; Adrian, V.; Matthew, K.G.; Team, T.G.B.; Joseph, P.P.; Dale, H.; Dan, C.; Peter, N.; et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* **2011**, *331*, 176–182.
54. Guus, S.; Alia, A.; Lora, A.; Mark, v.A.; Victor, d.B.; Lynda, H.; Michiel, H.; Borys, O.; Jacco, v.O.; Anna, T.; et al. Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *Web Semant. Sci. Serv. Agents World Wide Web* **2008**, *6*, 243–249.
55. Declerck, T. Towards a Linked Lexical Data Cloud based on OntoLex-Lemon. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
56. Tittel, S.; Bermúdez-Sabel, H.; Chiarcos, C. Using RDFa to Link Text and Dictionary Data for Medieval French. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.