A Semantic Model for Traditional Data Collection Questionnaires enabling Cultural Analysis

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

[1] Dublin City University, Adapt Centre. IE
[2] Austrian Academy of Sciences, Austrian Centre for Digital Humanities - AAS. AT

LDL2018 @ LREC2018. Miyazaki (JP) 12/05/2018

Presentation Outline

- Introduction
- Background
- Approach
- Semantic Modelling
- Semantic Uplifting
- Current work
- Future Directions

Introduction

- The research is motivated by
  The large amount of digitized traditional data, the open access policy adopted by institutions and the availability of technological solutions.
- Challenges
  However efficient utilization of existing resources by humans and machines is hindered by lack of semantics to understand and interpret the data.

  Interlinking of the data within or across external resources is often a challenge

Introduction

- Our focus is

  To facilitate effective opening up of the data

  To provide a semantic model for traditional data
  collection and analysis

  To provide a means of interlinking such data

Background

This research deals with traditional data including historical, socio-cultural, political, lexicographic data sets that are collected over an extended period.

- Language
  - German language, Bavarian dialect
- Geographic coverage
  - Austria, Czech Republic, Slovakia
  - Hungary & Northern Italy
- Data collection
  - Primary data collected using questionnaire
  - Secondary data collected using other sources

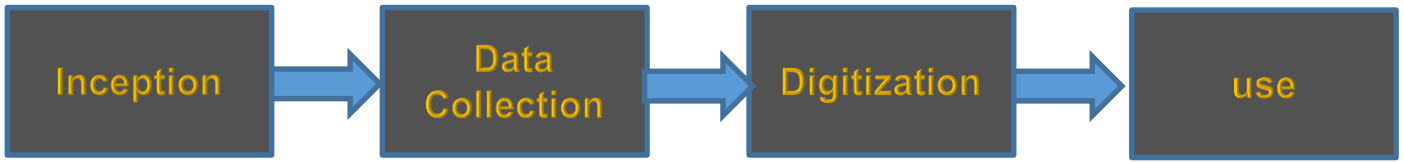
Background

- Domain
  - Lexicographic
  - Historical
  - Socio-cultural
- Time Period
  - 1913 - 1998
- Size
  - 120 questionnaires
  - 24,382 questions
  - 3.6 million paper slips
  - 11,157 individuals

Approach

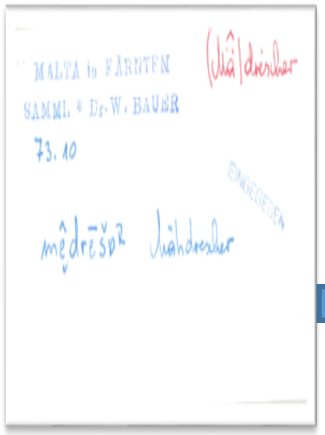
The approach we followed has two phases:

1. Schema Analysis
   - Serves as a means of understanding the data collection and processing steps during the project



Approach

- Schema Analysis
  - Shows different stages of processing and interpretations of the data over time.



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

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

Field	Type	Null	Key	Default	Extra
id	int(11)	NO	PRI	NULL	auto_increment
nummer	varchar(20)	NO	UNI	NULL	
titel	varchar(512)	NO		NULL	
schlagwoerter	varchar(1024)	YES		NULL	
erscheinungsjahr	int(11)	YES		NULL	
person_id	int(11)	YES		NULL	
originaldaten	text	YES		NULL	
anmerkung	text	YES		NULL	
freigabe	tinyint(1)	NO		NULL	
checkiet	tinyint(1)	NO		NULL	
wordleiste	tinyint(1)	NO		NULL	
druck	tinyint(1)	NO		NULL	
online	tinyint(1)	NO		NULL	
publiziert	tinyint(1)	NO		NULL	
fragebogen_typ_id	int(11)	YES	MUL	NULL	

Paper Slip | TUSTEP | TEI/XML | MySQL

Approach…

2. Domain Analysis

Domain analysis mainly answers the following questions

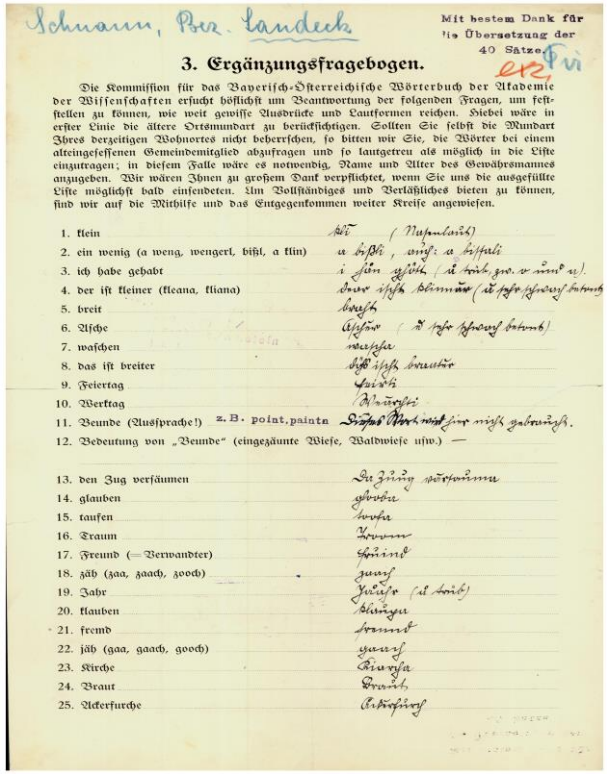
- The main purpose is collection and preservation of diversity of language and culture
- The source includes individuals, groups, organizations, written and verbal resources
- The domain covers lexicography, culture, history, economy and others
- The Scope covers mainly questionnaires and questions

Semantic Modelling

- We studied the questionnaires and their questions in detail. Their attributes, types, relations are identified.
- Are there reusable semantic models?
  - Yes
    - Schema.org
    - DublinCore
    - SKOS
    - Ontolex
  - No
    - Questionnaire
    - Questions, answers etc

Semantic Modelling

- Questionnaire Model



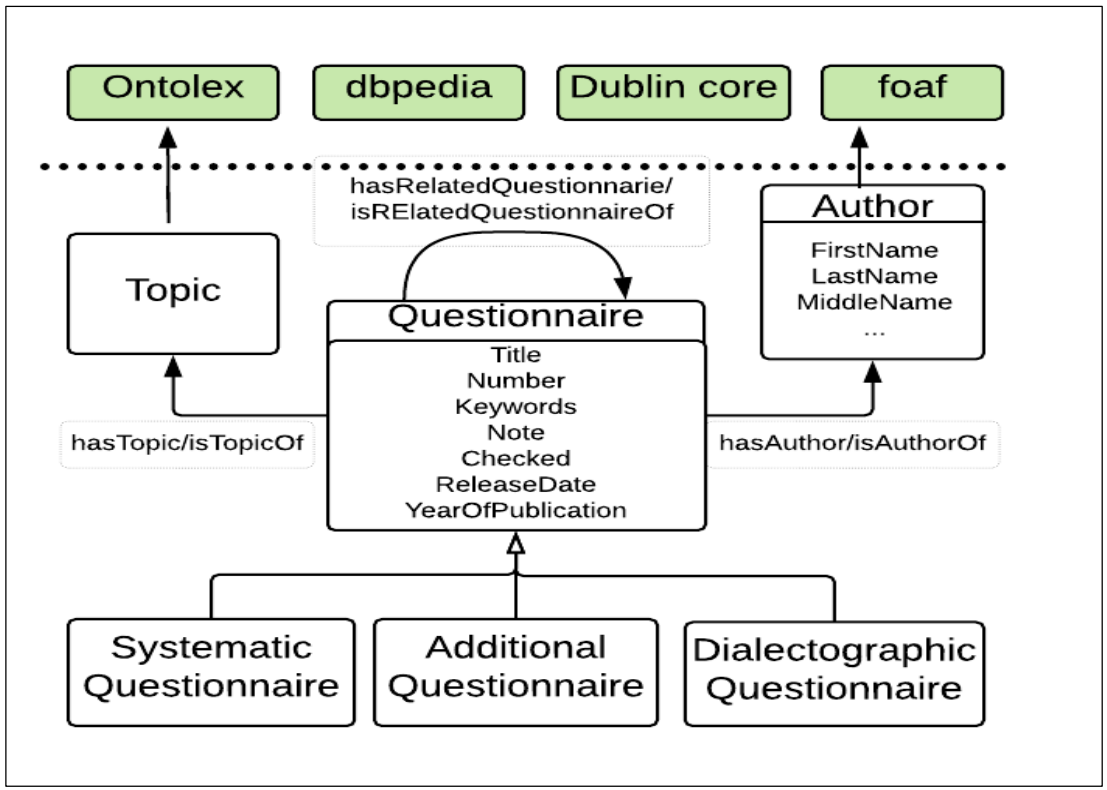
Schwarm, Post Landeck.

Mit bestem Dank für die Übersetzung der 40 Sätze.

3. Ergänzungsfragebogen.

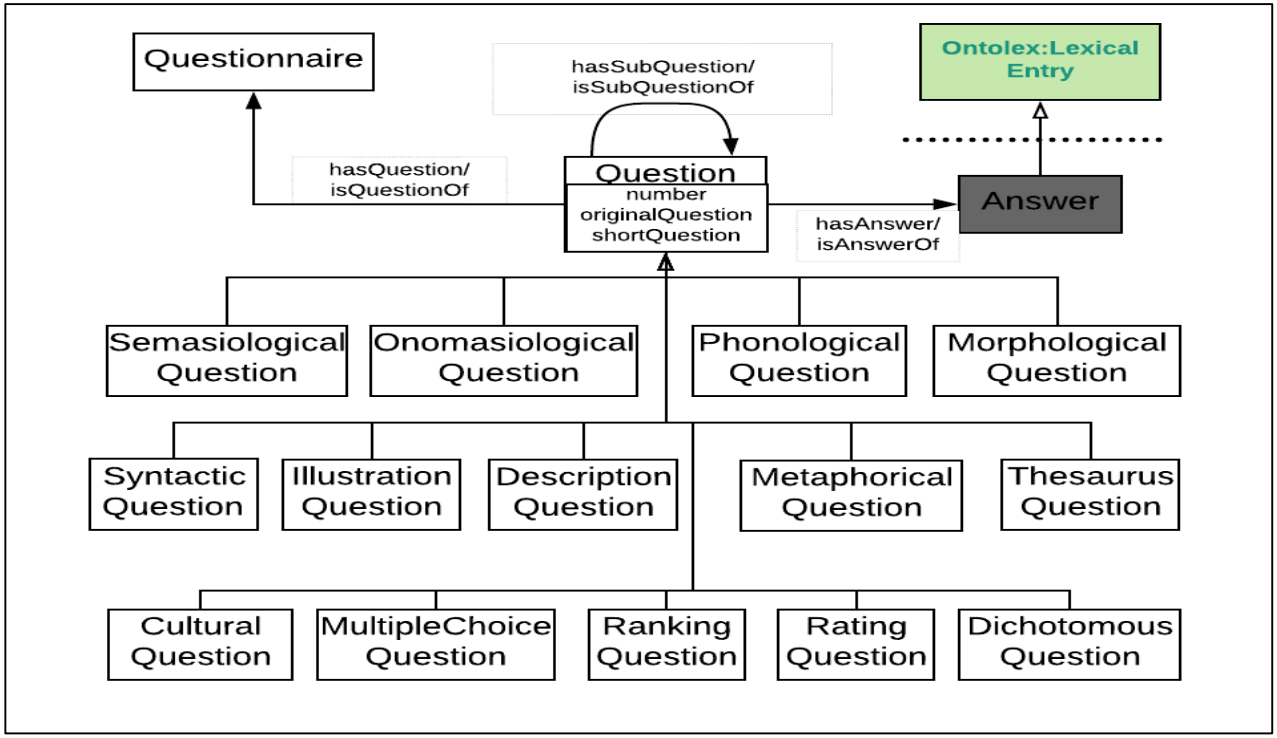
Die Kommission für das Bayerisch-Österreichische Wörterbuch der Akademie der Wissenschaften ersucht höflichst um Beantwortung der folgenden Fragen, um feststellen zu können, wie weit gewisse Ausdrücke und Lautformen reichen. Hiebei wäre in erster Linie die ältere Ortsmundart zu berücksichtigen. Sollten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

1. klein
2. ein wenig (a weng, wengerl, bißl, a klin)
3. ich habe gehabt
4. der ist kleiner (kleana, kliana)
5. breit
6. Asche
7. waschen
8. das ist breiter
9. Feiertag
10. Werktag
11. Beunde (Aussprache!) z. B. point, paintn
12. Bedeutung von „Beunde" (eingezäunte Wiese, Waldwiese usw.) —
13. den Zug versäumen
14. glauben
15. taufen
16. Traum
17. Freund (= Verwandter)
18. zäh (zaa, zaach, zooch)
19. Jahr
20. klauben
21. fremd
22. zäh (zaa, zaach, zooch)
23. Kirche
24. Braut
25. Ackerfurche



Semantic Modelling

- The Questions



What is next?

- Validating the model with the experts
- Creating the ontology
- Enriching the ontology by working with the experts
- Cleaning and repairing the data
- **Semantic enrichment of the data**

Semantic Uplifting

- **Relational 2 RDF Mapping**
- **R2RML**
- **RDF data generation**
- Support for API

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

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

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

Semantic Uplifting

- RDF serialization of the data.

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

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

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

Discussion

- Domain analysis is useful in understanding and preserving the original intent of the data and the data collection process.
- It enables the defining and describing entities and relationships which are difficult to understand without a proper description
- Schema analysis captures important entities and attributes and links that would not be identified otherwise. It further shows the evolution over time.

Discussion

- Complex entities and attributes are studied and understood.

- Despite the quality of data, some drawbacks are also identified
  - It requires domain experts
  - Takes a long time

Current word

- Further modelling of new entities including
  - Paper slip
  - Lemma
  - Source
  - Place
  - Person
- Interlinking with existing knowledgebase
  - Questionnaire –DBpedia
  - Locations – GIS sources
  - Lexical entries – existing dictionaries
  - Persons – person databases

Future work

- Exploration
- Semantic Search
- Semantic Bot

Questions