The Survey Ontology: Packaging Survey Research as Research Objects

Mario Scrocca, Damiano Scandolari, Gloria Re Calegari,
Ilaria Baroni, and Irene Celino

Cefriel – Politecnico di Milano
Viale Sarca 226, 20126 Milano, Italy
name.surname@cefriel.com

Abstract. Surveys are a common and well explored method to collect information from people. Still, the sharing and reuse of survey data present several challenges for survey researchers that need to be supported in packaging and harmonising different resources describing a survey study.

In this paper, we present the *survey ontology* that we designed to empower our CONEY toolkit for conversational surveys. Leveraging on Semantic Web technologies we aimed at building a solution to semantically annotate questions and answers at design time, and to easily elaborate and inter-link the collected data at analysis time. The *survey ontology* embraces the research object principles, and defines an open vocabulary to represent, annotate, and share a representation of the questionnaire structure and the gathered responses of a survey. We complement the discussion describing a complete survey research study carried out with CONEY and openly published as a research object.

Keywords: Survey Research · Survey Ontology · Conversational Survey · Research Object

1 Introduction

Questionnaire design [18] is a solid discipline that addresses the challenges of efficient and effective administration of surveys and of reliable data collection. In this field, the two main research methods – quantitative and qualitative – are often opposed and contrasted.

Qualitative research, usually carried out by means of interviews, is an observation method based on non-numerical information that favour human interaction; the challenges of qualitative design are related to the cognitive factors of respondents [15]. Quantitative research methods investigate phenomena via statistical, mathematical, or computational techniques applied to numerical data, thus supporting the collection and analysis of quantitative information.

In the analysis of collected answers, on the one hand, the responses to one survey must allow for a proper analysis of the investigated factors; on the other hand, it is common practice to re-submit the same survey – or variants of the same survey – either to different respondent groups (e.g., in A/B testing) or to the same users in different moments (e.g., in *ex ante* and *ex post* analyses), in order to compare and contrast results [18]. A semantic annotation of questions and answers can facilitate data analysis and easily identify correspondences between different surveys. Moreover, the use of a reference ontology allows for exporting both the survey structure and the collected answers as linked open data.

We conceived and implemented CONEY[1], a CONversational survEY toolkit, a system to design and administer questionnaires and analyze the collected results. CONEY implements a quantitative data collection, "disguised" as a qualitative research method by administering the questionnaire with a chat-like user experience [7]. CONEY allows the semantic annotation of questions and answers, being based on a common conceptual model.

This paper describes this conceptual model, the *survey ontology*; we designed it to empower CONEY, but we released it as an open vocabulary because it covers the main aspects of questionnaires, independently of the adopted survey tool. This ontology, relying on the research object principles, aims at addressing several challenges in the representation and packaging of survey data.

The remainder of the paper is organized as follows: Section 2 introduces the problem space and the challenges we address; Section 3 illustrates the *survey ontology*; Section 4 describes an end-to-end example of use and exploitation of CONEY to package a survey research object; final considerations are offered in Section 5.

2 Problem Space and Challenges

While questionnaire design is a well explored discipline with a wealth of different and robust methodologies, it is still an open issue how to simplify and support data and methods sharing. With our approach, indeed, we address the problem of enriching the entire pipeline of survey design, administration and result analysis, with the objective to make it easy to share and interlink data from both the survey structure, its collected answers and the analysis methods, in line with the Open Science principles [10] as well as the vision for Responsible Research and Innovation (RRI) [16]. In this section, we describe the relevant related work, and we discuss the main challenges considered in the definition of the *survey ontology*.

2.1 Related Work

The seminal paper from Bechhofer et al. [3] defines the concept of *Research Object* to enhance the publishing, share and reusability of research data through linked data. The *Research Object Suite of Ontologies* [4] focuses on the principles of identity, aggregation and provenance, defining a set of workflow-centric ontologies to represent research objects. The *survey ontology* extends the proposed

[1] https://coney.cefriel.com

approach by interpreting a complete survey research study as a research object, the survey procedure as a scientific *workflow* (*wfdesc* module) and the survey's collected answers as provenance traces of its *execution* (*wfprov* module). The RO-Crate [1] specification defines a lightweight approach to publish research objects. In this paper, we show an example of RO-Crate describing a survey study and exploiting the *survey-ontology* to represent the survey data.

The DDI-RDF Discovery Vocabulary [12], based on the Data Document Initiative (DDI) international standard, defines a vocabulary to represent survey data as linked data. DDI-RDF was considered for direct reuse but, because it is based on a document-centric approach, this option was discarded since we decided to adopt a workflow-centric approach for the *survey ontology*. Nevertheless, similar concepts are defined in the two ontologies and we aim at providing an alignment with the DDI-RDF vocabulary to enable the reuse of concepts not included in the *survey ontology*, e.g., the pattern to represent survey waves through DDI-RDF *Study* and *StudyGroup* concepts.

The GESIS - Leibniz Institute for the Social Sciences has been investigating the potentialities of adopting Semantic Web technologies for survey data in the social sciences. Gottron et al. in [12], discussed the problem of integrating different resources for a survey proposing a framework for a semantic data library for the social sciences. The authors focus on the integration of survey datasets and investigate the potentialities of adopting the *RDF Data Cube Vocabulary*[2] to standardise the representation of collected data and facilitate their integration. The importance of adopting controlled vocabularies to publish survey research, and the potentialities of inter-linking data from multiple surveys, is further discussed by Heling et al. in [14]. In this work, the authors describe a prototype pipeline exploiting an extension of the DDI-RDF vocabulary to represent in a single knowledge graph the survey data extracted from different structured and unstructured documents. On one hand, we considered the modelling decisions and the vocabularies adopted in this work for the definition of the *survey ontology*, on the other hand, the *survey ontology* can offer a valuable resource to enhance the representation in the knowledge graph of provenance information for survey data. The recent work from Bensmann et al. [5] defines a comprehensive vocabulary for the structured description of survey questions and their content dimensions. The vocabulary is very valuable for survey designers to enhance the research and reuse of relevant questions in previous studies. Questions modelled using the *survey ontology* can be extended using this vocabulary to enrich their semantic description.

Additional ontologies indexed in the LOV portal[3] and defining the concept of survey are out of scope: the *SemSur* vocabulary [9] is about literature reviews; the SIOC ontology [6] is about online communities, in which polls can be embedded in posts, but doesn't model questionnaires.

[2] RDF Data Cube Vocabulary, cf. `https://www.w3.org/TR/vocab-data-cube/`

[3] Linked Open Vocabularies, `https://lov.linkeddata.es/`

2.2 Open Issues

In this section, we provide an overview of the open issues in this problem space, which will guide also the explanation of our solution in the rest of the paper. The challenges that we address can be defined as follows:

C1 *Make the survey structure available as structured data*: to avoid the risk of "burying" the survey semantics in documents, we aim at providing a way to export a survey as a dataset by itself; as specified in [14], there is no established and specialized vocabulary for surveys, therefore we aim at providing a reference and reusable survey ontology.

C2 *Annotate questions with the respective investigated variables*: to make it possible to analyse survey results more easily, as well as to enable the comparison between different studies, we aim to allow the survey designer to annotate the questions with the variables or phenomena they are designed to investigate; those annotations can be re-used across different studies.

C3 *Annotate answers with their numerical coding*: to ease the result analysis, we aim to allow the survey designer to annotate also the questions' pre-defined answers with their numerical value for subsequent computation of mean, median, variance, etc.; for example, a 5-point Likert scale of answers like "strongly disagree/disagree/neutral/agree/strongly agree" can be annotated with numbers from 1 to 5.

C4 *Make the collected answers available as structure data*: to facilitate the subsequent analysis, we aim to allow for result export as a structured dataset as well, by employing the same survey ontology (cf. **C1**); this allows cross-linking between the survey structure and its results, as well as between different compiling campaigns of the same survey (e.g., in case of repeated administration of the same questionnaire to different groups of respondents or to the same group at different times).

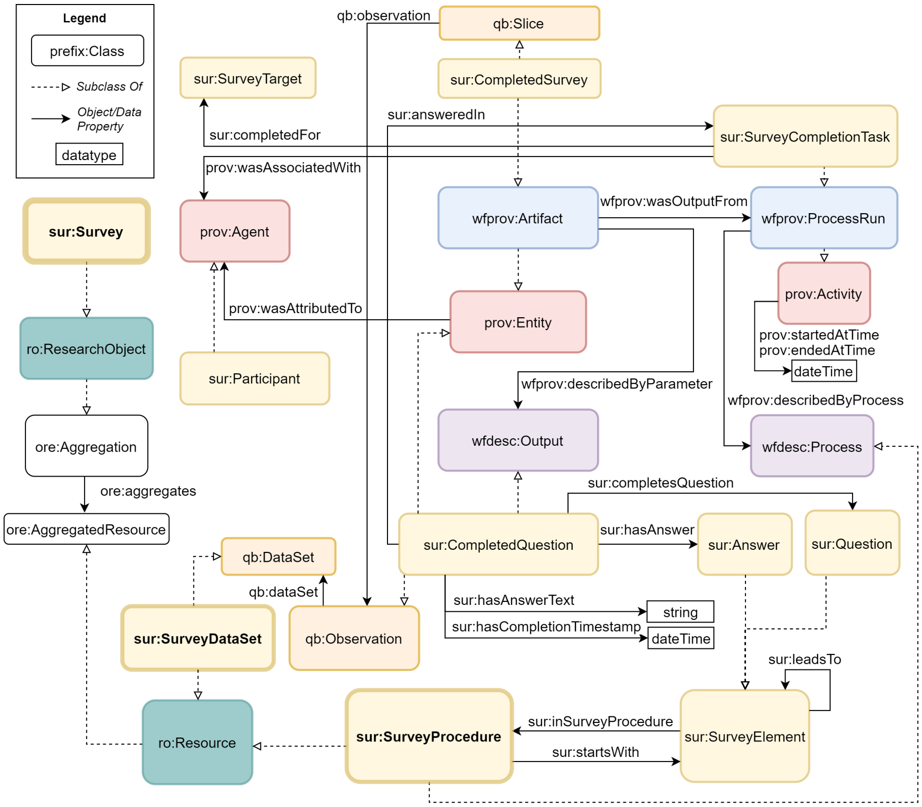
C5 *Keep provenance of answers*: to track the link between respondents and their answers, we aim at using provenance; this also helps in cross-study assessment, if the respondents are uniquely identified.

C6 *Share the survey methodology*: to foster repeatability and reproducibility of research, we aim at facilitating to share not only questions and collected answers, but also the scientific method behind it, like the hypothesis for correlation, causality and other interplay between the investigated variables, or the actual analysis processes and techniques, to pave the way for a full "research object" sharing.

3 The Survey Ontology

This section presents the *survey ontology* openly published at `https://w3id.org/survey-ontology`[4]. Challenges presented in Section 2.2 are additional requirements for the design phase, and we discuss how they were addressed.

[4] The endpoint offers content negotiation and we generated the HTML documentation using Widoco [11]. In the related repository (cf. `https://github.com/cefriel/survey-ontology`), we also provide the related OOPS [17] evaluation.



sur w3id.org/survey-ontology#, *ore* http://www.openarchives.org/ore/terms/, *qb* http://purl.org/linked-data/cube#, *prov* http://www.w3.org/ns/prov#, *ro* http://purl.org/wf4ever/ro#, *wfprov* http://purl.org/wf4ever/wfprov#, *wfdesc* http://purl.org/wf4ever/wfdesc#

Fig. 1: The Survey Ontology

In the development of this ontology, we followed ontology design principles [13] and best practices. For this reason, we designed our ontology in a generic still extensible way to cover any kind of survey, reducing the specificity of its application to conversational surveys. In the ontology design, we adopted the Linked Open Terms (LOT) methodology [20], and we devised a set of use cases and related competency questions made available in the ontology repository[5].

The core concepts of the ontology (cf. Figure 1) are designed to model the basic elements composing the structure of the survey (**C1**). We modelled a generic survey, with the additional inclusion of characteristics of a conversational survey approach. Hereafter, we give an overview of the main concepts; for space reasons, we do not explain the entire ontology.

[5] Survey ontology wiki, cf. `https://github.com/cefriel/survey-ontology/wiki`.

The main concept is the *SurveyElement*, representing each building block of a survey. In the first version of the ontology, we identified three main subtypes of blocks: *Question*, *Answer* and *Talk*. The connection between the different blocks is defined through the property *leadsTo* connecting subsequent *SurveyElement*s.

A *Question* block is described by the textual description of the question and defines the typology of interaction required to the user. We identified two main subclasses of *Question*: *OpenQuestion*, accepting an arbitrary free-text answer from the user, and *ClosedQuestion*, offering a set of options to choose from.

To annotate questions with investigated variables (**C2**), we defined two specific concepts: the *ObservableVariable* class, which describes the variable measured by the question (e.g., feedback on an event catering), and the *LatentVariable* class, which identifies the indirectly measured variable (e.g., overall appreciation for the event). In general, each variable individual can be defined as an instance of these classes; the use of specific vocabularies can help other researchers in searching, in a knowledge graph containing published surveys, the set of questions addressing the same latent/observable variables and asked in other related studies.

Answer blocks have two main subclasses: *OpenAnswer* blocks that can be associated only to *OpenQuestion*s and describe the typology of input expected by the user, and *ClosedAnswer* blocks associated to *ClosedQuestion* blocks and characterizing the available options for each question. To annotate answers with their numerical coding (**C3**), we added the *hasValue* data property to associate a numeric value to each answer instance Finally, a *Talk* block represents a textual message within the survey, it can be characterized by a simple text or a link to an external resource.

To address the remaining challenges, we decided to frame the newly introduced concepts in the context of the *Research Object Suite of Ontologies* guaranteeing a solid model to structure research data (**C4**), handling provenance with PROV-O[6] (**C5**) and fostering open science principles (**C6**).

The main concept is the *Survey* class that is defined as a subclass of *ResearchObject* (**C6**). Intuitively, a *ResearchObject* is an aggregation of *Resource*s describing a scientific investigation. A *Survey* aggregates two main resources identified in our ontology: a *SurveyProcedure*, describing the survey structure, and a *SurveyDataSet*, containing collected answers. Moreover, being a *Research-Object*, a *Survey* can aggregate other additional resources, representing study hypotheses, investigated variables, models produced from the result analysis, related publications, etc.

A *SurveyProcedure* describes the structure of the survey and it is defined as a subclass of *Process* since it represents the "workflow" adopted in the research study to collect answers. A *SurveyProcedure* is connected to all the *SurveyElement*s composing the survey, the first block is identified through the *startsWith* object property (**C1**).

The *SurveyCompletionTask* class inherits from *ProcessRun* and refers to the *Activity* of executing a *SurveyProcedure* and, hence, completing the survey. A

[6] Provenance Ontology, cf. `https://www.w3.org/TR/prov-o/`

CompletedSurvey is the *Artifact* generated as a result of a *SurveyCompletionTask*, and it is described by a set of *CompletedQuestions* representing the answers given by a specific user in specific survey completion. As such, also all the information collected from respondents during the survey completion is made available as structured data (**C4**) as well as interlinked to the survey structure.

To handle provenance of collected answers (**C5**), we defined the *Participant* class identifying the PROV-O *Agents* associated with the *SurveyCompletionTask*, the resulting *CompletedSurvey* and the related *CompletedQuestions*. It is important to point out that a single *Participant* may compile the survey multiple times, thus having multiple *SurveyCompletionTasks* associated.

A *SurveyDataSet* is a *DataSet* as defined in the *RDF Data Cube vocabulary* and collects all the *CompletedQuestions* (subclass of *Observation*) for a *Survey*.

An RDF representation of survey data adopting the *survey ontology* defines an integrated and structured representation of both the survey procedure and the collected answers (**C1**, **C4**) that can be shared as a *research object* aggregating all the resources for the considered study (**C6**). An example of a knowledge graph using our *survey ontology* is depicted in Figure 2, which represents a portion of the TESS dataset (discussed in Section 4) to showcase the integrated representation. The red circle represents the TESS *SurveyProcedure* connected to the set of *SurveyCompletionTasks*, each one representing a collected compilation of the survey. The *SurveyProcedure* is also connected to the survey elements represented by the chain of circles starting with the *startsWith* relation and continuing by means of the *leadsTo* property. The first *ClosedQuestion* of the chain is shown, together with the three *ClosedAnswers* associated (in grey). We

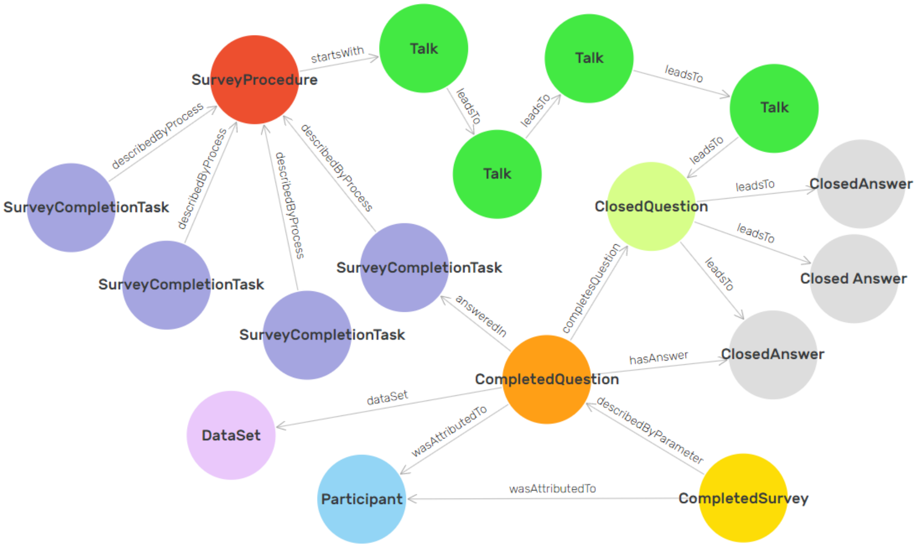


Fig. 2: A portion of the TESS survey knowledge graph.

expanded one *CompletedQuestion* to showcase how it is bound to the answer chosen by the *Participant*, to the *SurveyCompletionTask*, to the *CompletedSurvey* and to the overall *SurveyDataSet*.

In the design of the ontology, we chose to limit the use of logical axioms and we defined a set of SHACL shapes[7] for validating data represented using the *survey ontology*.

4 The TESS Network Motivation Survey

In this section, we describe how we exploited the *survey ontology* to address challenges described in Section 2.2 by packaging and sharing a research object describing a survey study for users involved in a citizen science campaign.

The goal of the study [8], conducted within the ongoing ACTION H2020 project[8], is to analyse the motivation to participate, of a specific citizen science community, in the effort of fighting light pollution: the network of around 120 hosts of the TESS photometers[9].

The designed survey investigates 14 latent variables related to motivation and data collection in citizen science. We kept the survey balanced by selecting two questions for each variable and we customized the formulation of the questions in order to make them more specific to the TESS photometer context. All the questions were designed to have closed answers and have been annotated with the respective latent variable (**C2**); the answers are associated with both a qualitative value, a textual label to be displayed in the chat, and a quantitative value, the numerical coding for results analysis (**C3**).

We collected answers from 79 volunteers, corresponding to the 65% of our target users. Analysing the results of the survey, we computed the mean value of the answers related to each motivating factor, and the correlation of each factor with the global motivation to participate. The values of the answers range from 1 to 5 by construction, and the value of global motivation has been collected by asking directly the compilers their perceived level of motivation to participate in such a citizen science initiative. As an example, we discovered that a high motivation correlates with a strong willingness to participate to make data more accessible and to raise public awareness on the light pollution problem (cf. *universalism* latent variable).

The survey designed for the TESS network is not limited to this specific initiative, but it can be used by different survey researchers to study the motivations of different citizen science communities. To promote the reuse of the presented approach, we exploited the *survey ontology* and the CONEY Toolkit to export the relevant resources and we defined a research object for the TESS network motivation survey. A comprehensive RO-Crate[10] is made openly avail-

[7] `https://cefriel.github.io/survey-ontology/ontology/sur_shapes.ttl`

[8] ACTION (pArticipatory sCience Toolkit agaInst pollutiON) project, cf. `https://actionproject.eu/`

[9] TESS Photometer Network, cf. `https://tess.stars4all.eu/`

[10] Packaged using Describo `https://uts-eresearch.github.io/describo/`

able on Zenodo [19], including the representation of the survey structure (**C1**), the collected answers (**C4**) with provenance information (**C5**), the script and results of the analysis, and related publications (**C6**). The adoption of the *survey ontology* helps in identifying links among the resources in the research object: the survey dataset of collected answers is bound to the *SurveyElement*s in the survey structure, and the results of the analysis refer to the *LatentVariable*s used to annotate questions in the survey.

5 Conclusions

In this paper, we presented the *survey ontology*, which is the conceptual data model behind our CONEY toolkit, but it is also a generic and comprehensive open vocabulary to describe any kind of survey; this appeared as a missing element in the panorama of available ontologies [14], and some concepts of our survey ontology are strongly related to the ones of relevant models with complementary scope that we reused and interlinked (like PROV-O, Data Cube, and the Research Objects suite of ontologies). We believe that both the toolkit and the ontology are interesting for the Semantic Web Community – as well as anybody who wants to create a survey – both to support a survey-based investigation and to allow for data collection, representation and analysis based on Semantic Web technologies and Open Science principles. In this context, the *survey ontology* can foster the packaging and share as research objects of the survey structure, the investigated variables and the results of the analysis. To support our claims, we published and described a comprehensive research object for a survey study performed using CONEY and exploiting the *survey ontology* to describe the relevant resources.

As future work, we plan to investigate the reuse of the Research Variable Ontology [2] to represent models obtained from data analysis and to extend CONEY to automate the generation of research objects compliant with the RO-Crate specification.

Acknowledgments The presented research was partially supported by the H2020 ACTION project (Grant Agreement 824603), co-funded by the European Commission.

References

1. A lightweight approach to research object data packaging. Zenodo (Jun 2019). https://doi.org/10.5281/zenodo.3250687, Abstract accepted for talk at Bioinformatics Open Source Conference (BOSC2019).
2. Bandara, M., Behnaz, A., Rabhi, F.A.: RVO - The Research Variable Ontology. In: European Semantic Web Conference. pp. 412–426. Springer (2019)
3. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C.: Why linked data is not enough for scientists. Future Generation Computer Systems **29**(2), 599–611 (2013). https://doi.org/10.1016/j.future.2011.08.004

4. Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J.M., Bechhofer, S., Klyne, G., Goble, C.: Using a suite of ontologies for preserving workflow-centric research objects. Journal of Web Semantics **32**, 16–42 (2015). https://doi.org/10.1016/j.websem.2015.01.003
5. Bensmann, F., Papenmeier, A., Kern, D., Zapilko, B., Dietze, S.: Semantic annotation, representation and linking of survey data. In: International Conference on Semantic Systems. pp. 53–69. Springer, Cham (2020)
6. Breslin, J.G., Decker, S., Harth, A., Bojars, U.: Sioc: an approach to connect web-based communities. International Journal of Web Based Communities **2**(2), 133–142 (2006)
7. Celino, I., Re Calegari, G.: Submitting surveys via a conversational interface: an evaluation of user acceptance and approach effectiveness. International Journal of Human-Computer Studies **139**, 102410 (2020)
8. Celino, I., Re Calegari, G., Scrocca, M., Zamorano, J., González Guardia, E.: Participant motivation to engage in a citizen science campaign: the case of the TESS network. Journal of Science Communication (JCOM) p. (to appear) (2021), Third International ECSA Conference
9. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Semsur: a core ontology for the semantic representation of research findings. Procedia Computer Science **137**, 151–162 (2018)
10. Fecher, B., Friesike, S.: Open science: one term, five schools of thought. In: Opening science, pp. 17–47. Springer (2014)
11. Garijo, D., et al.: WIDOCO 1.4.13: Linking evaluation in documentation (2020). https://doi.org/10.5281/zenodo.3605675
12. Grotton, T., Hachenberg, C., Harth, A., Zapilko, B.: Towards a semantic data library for the social sciences. In: International Workshop on Semantic Digital Archives. vol. 801, pp. 48–59. DEU (2011)
13. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? International journal of human-computer studies **43**(5-6), 907–928 (1995)
14. Heling, L., Bensmann, F., Zapilko, B., Acosta, M., Sure-Vetter, Y.: Building Knowledge Graphs from Survey Data: A Use Case in the Social Sciences. In: Knowledge Graph Building Workshop, co-located with the Extended Semantic Web Conference 2019 (2019)
15. Ongena, Y.P., Dijkstra, W.: A model of cognitive processes and conversational principles in survey interview interaction. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition **21**(2), 145–163 (2007)
16. Owen, R., Macnaghten, P., Stilgoe, J.: Responsible research and innovation: From science in society to science for society, with society. Science and public policy **39**(6), 751–760 (2012)
17. Poveda-Villalón, M., Gomez-Perez, A., Suárez-Figueroa, M.C.: OOPS! (OntOlogy Pitfall Scanner!): An on-line tool for ontology evaluation. International Journal on Semantic Web and Information Systems **10**, 7–34 (04 2014). https://doi.org/10.4018/ijswis.2014040102
18. Saris, W.E., Gallhofer, I.N.: Design, evaluation, and analysis of questionnaires for survey research. John Wiley & Sons (2014)
19. Scandolari, D., Calegari, G.R., Scrocca, M., Celino, I.: Tess network motivation survey. https://doi.org/10.5281/zenodo.5140351
20. Villalón, M.P., Izquierdo, A.F., Castro, R.G.: Linked Open Terms (LOT) Methodology (2019). https://doi.org/10.5281/zenodo.2539305