**JOINT FINAL REPORT TO THE PROJECT (public part)**

**GEMEINSAMER ABSCHLUSSBERICHT ZUM PROJEKT (öffentlicher Teil)**

# A pilot study for "Linked Open Research Data" (LORDpilot): a LOD-based Concept Registry for social science research data

# Pilotstudie für „Linked Open Research Data“ (LORDpilot): eine LOD basierte Concept Registry für sozialwissenschaftliche Forschungsdaten

**Authors in alphabetic order:**

Dr. Andreas Daniel, Dr. Jan Goebel, Dr. Dagmar Kern, Daniel Klein,
Antonia May, Fakhri Momeni, Jana Nebelin, Claudia Saalbach,
Dr. Pascal Siegers, Knut Wenzig & Dr. Benjamin Zapilko

Berlin, Hannover & Köln: November 2023

## 1. General Information

DFG reference number: (DA 2545/1-1 | GO 2651/5-1 | KE 2311/3-1 | SI 2022/4-1)

Project number: 464413245

Project title:

DE: „Pilotstudie für „Linked Open Research Data“ (LORDpilot): eine LOD-basierte Concept Registry für sozialwissenschaftliche Forschungsdaten“

EN: “A pilot study for "Linked Open Research Data" (LORDpilot): a LOD-based Concept Registry for social science research data”

Names of the applicants:

Dr. Andreas Daniel[1], Dr. Jan Goebel[2], Dr. Dagmar Kern[3], and Dr. Pascal Siegers[3]

Official address(es):

[1] Deutsches Zentrum für Hochschul- und Wissenschaftsforschung, Lange Laube 12, 30159 Hannover

[2] DIW Berlin, Mohrenstraße 58, 10117 Berlin

[3] GESIS - Leibniz-Institut für Sozialwissenschaften, Unter Sachsenhausen 6-8, 50667 Köln

Name(s) of the co-applicants: Knut Wenzig, Dr. Benjamin Zapilko

Reporting period (entire funding period): 05/2022-09/2023

## 2 Summary / Zusammenfassung

**DE:** Die Nachnutzung von Forschungsdaten ist ein wichtiger Bestandteil der Forschungspraxis in den Sozial- und Wirtschaftswissenschaften. Um geeignete Daten zu finden, brauchen Forschende funktionierende Suchangebote. Eine übergreifende Suche nach Daten wird jedoch durch eine uneinheitliche oder fehlende semantische Erschließung erschwert, weil verschiedene Erhebungsprogramme jeweils eigene Terminologien für die Dokumentation verwenden. Meist fehlt auch eine Verknüpfung der gemessenen theoretischen Konzepte mit den Variablen. Aus Sicht der Nutzenden behindert die Fragmentierung in der Datendokumentation die Datensuche und schränkt deshalb das Forschungspotential existierender Bestände ein. Die Herausforderung liegt deshalb in der konzeptorientierten Erschließung von Daten. Weil eine semantische Modellbildung für die inhaltliche Erschließung bislang fehlt, werden ein Prozess und eine Technologie für eine einheitliche, semantische Indexierung der Forschungsdaten benötigt. Die LORD-Infrastruktur soll diese Lücke schließen.

Ziel des Projektes ‚LORDpilot‘ war es, die Machbarkeit einer Concept Registry für die Sozialwissenschaften zu prüfen. Dazu wurden im Pilotprojekt ein Datenmodell und eine benutzerfreundliche Eingabemaske (AnnoTool) entwickelt, mit deren Hilfe für eine Auswahl von Messinstrumenten aus drei großen Umfragen (ALLBUS, nacaps, SOEP) Fragen bzw. Variablen mit theoretischen Konzepten verknüpft (d.h. annotiert) wurden. Für die technische Umsetzung wurden Standards des Semantic Web verwendet. Durch die Verknüpfung der Konzepte mit

Deskriptoren aus dem SCOS-konformen „Thesaurus Sozialwissenschaften“ (TheSoz) wird die Suche in der Konzeptdatenbank unterstützt und das Konzeptvokabular direkt in die Linked Open Data (LOD) Cloud integriert. Die Verknüpfungen wurden in Form von RDF-Triples erstellt und in einem Triple-Store mit SPARQL-Endpunkt zugänglich gemacht.

Für die Evaluation des Verfahrens wurden die ausgewählten Messinstrumente der drei Befragungen von jedem der beteiligten Projektpartner annotiert (d.h. Fragen und Variablen mit Konzepten beschrieben) und anschließend die Passung von Frage und Konzept von Fachexperten bewertet. Die Auswertung dieser Testannotationen zeigt, dass (1) die Annotationen verschiedener Annotatoren eine hohe Übereinstimmung aufweisen, (2) die Konzepte von den Fachexperten überwiegend als zur Messintention passend bewertet werden und (3) über die vergebenen Konzepte konzeptionelle Zusammenhänge über die Datensätze hinweg sichtbar werden. Allerdings zeigt die Auswertung auch, dass die Verwendung marginal unterschiedlicher Konzeptbegriffe irrelevante Heterogenität im Konzeptvokabular erzeugt.

Die Pilotstudie hat gezeigt, dass die im Antrag skizzierte Infrastruktur realisierbar ist, wenn die Redundanz im Konzeptvokabular begrenzt wird, z.B. indem durch algorithmische Unterstützung bei der Annotation passende Begrifflichkeiten vorgeschlagen werden.

**EN:** Reusing research data is an important part of research practice in the social and economic sciences. To find suitable data, researchers need functional search options. However, a comprehensive search for data is hampered by inconsistent or missing semantic indexing because different survey programs use their own terminology for documentation. In most cases, there is no link between the measured theoretical concepts and the variables. From the user's perspective, the fragmentation of data documentation hampers data retrieval and thus limits the research potential of existing data. The challenge, therefore, lies in the concept-oriented indexing of data. Since semantic modelling for content indexing is still lacking, a process and a technology for a uniform semantic indexing of research data are needed. The LORD infrastructure aims to close this gap.

The LORDpilot project aimed to test the feasibility of a concept registry for the social sciences. To this end, the pilot project developed a data model and a user-friendly interface to link (i.e., annotate) questions and variables with theoretical concepts for a selection of measurement instruments from three large surveys (ALLBUS, Nacaps, SOEP). We used Semantic Web standards for the technical implementation. By linking the concepts with descriptors from the SKOS-compliant "Thesaurus Social Sciences" (TheSoz), the search in the concept database is supported, and the concept vocabulary is linked to the Linked Open Data (LOD) Cloud. The links were created as RDF triples and made available in a triple store with a SPARQL endpoint.

To evaluate our approach, selected measurement instruments of the three surveys were annotated (i.e., questions and variables were described with concepts) by each of the project

partners involved, and then the fit between the measurement and the concept was assessed by domain experts. The evaluation of these test annotations shows that (1) the annotations of different annotators show a high degree of agreement, (2) the topical experts predominantly rate the concepts as matching the measurement intention, and (3) conceptual correlations across the data sets become visible via the assigned concepts. However, the analysis also shows non-substantive heterogeneity in the concept vocabulary across annotators.

The pilot study has shown that the infrastructure outlined in the application is feasible if the redundancy in the concept vocabulary is limited, e.g. by suggesting appropriate existing terms through algorithmic support during annotation.

## 3 Progress Report

Reuse of research data is crucial in the social and economic sciences, but finding relevant data is hampered by fragmented documentation and a lack of theoretical concepts in semantic indexing. The planned LORD infrastructure aims to develop a user-driven concept registry of social science and economics concepts for semantic data indexing. The aim is to make research data more discoverable and comparable for humans and machines by incorporating theoretical concepts into data documentation and linking them to metadata at the measurement level as it is required by the FAIR principles.

The more specific objective of the LORDpilot project was to assess the feasibility of user-driven concept-oriented indexing of questions and variables in social and economic research data.

To evaluate the basic components of the concept registry, we prototyped its main components: (1) a data model, (2) an annotation application for linking concepts to measures (AnnoTool), (3) and a triple store for storing concepts, metadata, and the links between them. The analysis of a test annotation of selected measures from three large German survey programmes. This enabled us to better understand the technologies and processes required for the successful implementation and operation of the infrastructure. We also identified potential obstacles and outlined possible solutions for the implementation of the infrastructure.

### Work Package 1: Research and Planning

The first work package of the project identified measurements (e.g., questions and variables) on comparable topics (e.g., concepts) in the three survey programmes (ALLBUS, SOEP, Nacaps) that were then used for our test annotation (see Work Package 4). The selection followed four criteria because we aimed for a large diversity of measures in terms of: (i) concept granularity and degree of concept standardisation, (ii) types of questions (e.g., question formats), (iii) target objects and time frames, and (iv) diversity of research areas using the same measures. The final selection includes concepts with highly standardised measurements (e.g., Big-Five personality traits, life satisfaction) but also topics with a high diversity of measures (e.g., xenophobia). We also decided to include questions on topics from adjacent research disciplines (e.g., subjective health indicators), and measures with different attitude objects (e.g., left-right self- and party placement) and time frames (e.g., retrospective and prospective perception of the economic situation). As a result, the LORDpilot corpus for test annotation comprises a total of 800 questions covering a wide range of topics.

To test the semantic indexing, the metadata of the measurements, consisting of question and item texts and their corresponding variable information, were collected, harmonised and imported into the AnnoTool for annotation (see work package 4). As metadata were stored in different formats in the different survey programmes the harmonisation of the metadata required considerable effort before the test annotation could begin.

### Work package 2: Data model

The term 'concepts' in the social sciences and humanities can refer to very different elements of reality: from the larger social framework (e.g., democracy), to elements or attributes of that framework (e.g., electoral participation), to latent social constructs (e.g., satisfaction with democracy). In practice, concepts are used for both theoretical reasoning and empirical investigation, the latter involving the specification of the relevant attributes of a concept and of a measurement for data collection. Taking this into account we have developed a flexible data model for the project that acknowledges the internal complexity of concepts and the relationships between concepts with the dimensions: concept, variable, and measurement (see Figure 1). The concept dimension deals with the complexity of concepts and the different descriptors used across fields and theoretical traditions. It captures the diversity of concepts and the relationships between them by allowing for (i) hierarchical relationships, (ii) relationships across research fields, and (iii) the collection of a variety of attributes. It therefore acts as an anchor to link studies, serving the wider purpose of increasing the retrievability and reusability of collected data. The dimensions of measurement and variables relate to the manifestation of concepts in the processes of data collection, processing, and archiving. To ensure integration with existing data management systems at the data-holding institutions, these dimensions also include the metadata, e.g., variable names and labels, question wordings and item texts. As the measurement dimension does not necessarily consist of questions and items, the model can be adapted to other data types, such as qualitative data and digital behavioural data.

According to the best practices of Semantic Web and Linked Data, the data model relies on definitions from existing standards and established vocabularies. Only in cases where no semantically adequate properties and classes exist, we add properties and classes defined in our own namespace lord: //data.gesis.org/lord/.

The model (Figure 1) captures the main classes like measurement, question, variable, and concept as well as the relations between them. The 'question' is a special case of a measurement when working with survey data. Other types of measurements could also be used. For most of the classes and properties, equivalent properties and classes are reused from schema.org[1], the Simple Knowledge Organization System (SKOS)[2], and the DDI-RDF Discovery Vocabulary (Disco)[3].

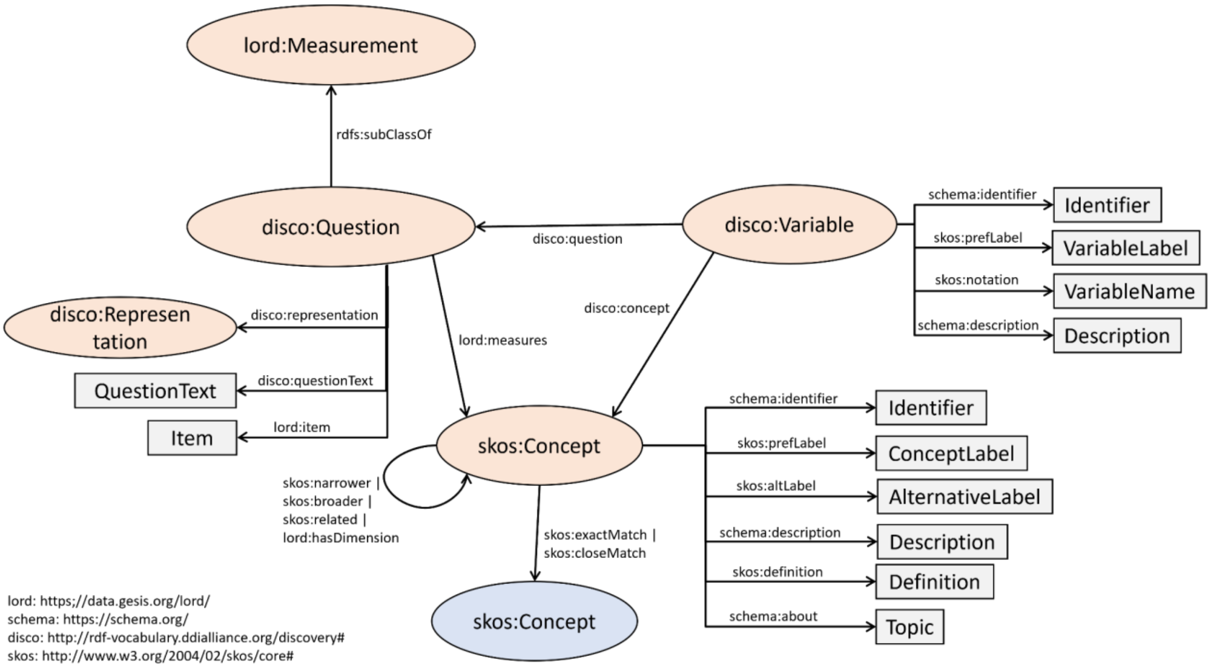

Figure 1: Overview of the LORD data model

Work package 3: Technical Environment

To prepare the test annotation of the 800 questions selected in Work package 1, we developed, using the Angular Framework, a web-application called AnnoTool that supports researchers in assigning variables with concept descriptors.

As an input file, the AnnoTool requests a CSV file in a predefined format containing all the necessary metadata of a question and/or variable in one line. After loading this file, the user interface displays the question and item texts, scales, variable names and labels. The annotator is then asked to first select one or more TheSoz terms that describe the topic or theoretical concept that the question is supposed to measure. To do this, we utilised the TheSoz API[4], which allows searching the thesaurus directly. In case the TheSoz descriptor is not sufficiently precise, the annotator can add terms as additional concepts that have been added by other users before. If there is no match at all, the annotator can add a new descriptor. To limit duplicates of descriptors (e.g., "left-right self-placement" vs "right-left self-placement"), all concept

[1] https://schema.org/, last access on November 16 2023.

[2] https://www.w3.org/TR/skos-reference/, last access on November 16 2023.

[3] https://rdf-vocabulary.ddialliance.org/discovery.html, last access on November 16 2023.

[4] https://lod.gesis.org/thesoz/de/sparql/thesoz/de, last access on November 16 2023.

terms added by all annotators are stored in a database and are available for selection immediately after being added. The decision to first select terms from TheSoz before free keywords can be amended is based on previous experience: First, the use of a controlled vocabulary, as provided by TheSoz, leads to a limitation of the heterogeneity of concept terms with the same meaning. This also concerns the control of spelling errors. Second, by using the TheSoz a connection to an existing linked open data infrastructure is created, which both relates the concept terms within the thesaurus and includes different language versions. Third, TheSoz is mapped to STW, which further extends the linkage to open data.

As the AnnoTool was only designed for the pilot study to assess the feasibility of the concept registry, the follow-up version of the tool must first check whether the desired descriptor is already registered in the concept registry. If this is not the case, the annotators should be able to suggest new descriptors and add them to the concept registry. In its current version, the AnnoTool does not allow to manually create links between concepts (e.g., personality has the dimension openness).

Eventually, the data from the AnnoTool was exported and transformed into RDF triples according to the data model defined in WP 2. The generated triples were imported into a Virtuoso triple store. The triple store provides a SPARQL endpoint[5] which allows for querying the concept registry, e.g., querying for certain concepts related to a particular variable, etc.

## Work package 4: Identifying and Linking Concepts

In this work package, we used the AnnoTool developed in WP 3 to select and link concepts to the 800 measurements selected in WP1. We used the test annotation to (i) assess whether the links between concepts and measurement are meaningful, (ii) to understand the processes of concept annotations and possible problems from the annotators perspective, (ii) identifying possible technical hurdles during annotation, and (iv) testing the integration of existing thesauri. The AnnoTool imports the harmonised metadata input file generated in WP1.

The four annotation teams selected 349 different concept terms to annotate the 800 measurements. Despite the large diversity in selected topics and question formats, the annotations did substantively overlap across annotators and datasets. In general, annotators selected higher level topic terms in addition to more specific terms describing the measurement, creating therewith links between the lower-level concepts and even bridging between different topic domains (disability, for example, is mentioned with health **and** discrimination). TheSoz thereby served as a useful reservoir for topic terms that facilitated bridging and linking between more specific concepts. Thus, the annotations created a network structure between concepts that could serve to design a visual search interface for users.

[5] https://sora.gesis.org/sparql, last access on November 16 2023.

To evaluate the semantic quality of the annotations, we asked domain experts to rate the fit between concepts and measurement intentions. The 11 experts rated 8,250 concept annotations for 708 variables/questions with an average score of 3.94 on a five-point Likert scale with 1 being "the concept does not fit the measurement at all" and 5 being "the concept fits the measurement very well". The terms added manually to the concept vocabulary scored even higher with a mean of 4.35. Overall, 93.94 percent of all measurements were annotated with at least one concept that was rated 5. This means: The annotation has successfully linked survey data with concepts with very good fit by our experts.

Our test annotation also revealed challenges. We observed that terms manually added to the concept vocabulary sometimes vary only slightly, which results in non-meaningful heterogeneity in the concept vocabulary. For example, one annotator used the concept "Big Five – Openness" and another only "Openness". This leads to inflating the concept registry and blurs similarities between measurements. The future project phase should include a system to avoid the emergence of "non-substantive" differences (see WP6).

Work package 5: Publications, Documentation and Networking

We focused our efforts on engaging with the community via networking events and international conference presentations. We presented the (preliminary) results at four international conferences (see section 4 for details and references to the presentations that have all been uploaded to zenodo). To share information with the community of data providers we also set up a website with basic information about the project.[6]

The feedback underscored the need to integrate concept terms into data documentation. The community welcomed the open and user-driven approach but also expressed doubt about the motivation of primary investigators to put additional efforts into data documentation. Several suggestions were made especially on creating links to repository software like dataverse or to software used for online data collection (e.g., LimeSurvey) as this would allow for annotation when the questionnaire is designed. Feedback also hinted at existing technologies for the management of vocabularies and ontologies, like for example WissKi[7] and Antelope[8].

Furthermore we engaged in exchanges with experts from national (RatSWD) and international (CESSDA) data infrastructures. One objective was to attract partners for the next project phase. But the majority of the European data archives still do not produce granular metadata. One notable exception is the Dutch data infrastructure consortium ODISSEI that expressed its

[6] https://www.diw.de/de/diw_01.c.862891.de/projekte/linked_open_research_data_for_social_science_pilot_study__lordpilot.html, last access on November 16 2023.

[7] https://wiss-ki.eu/de, last access on November 16 2023.

[8] https://nfdi4culture.de/de/dienste/details/annotation-terminology-lookup-and-personalization-antelope.html, last access on November 16 2023.

interest in a possible collaborative use of a concept registry. Similarly, some German RDCs consider the use of a concept registry like for example the LIfBi.

Work package 6: Concept development & follow-up application

In Work Package 6, we summarised the findings of WP1 to WP4 and the feedback received in WP5 to outline possible solutions to the problems identified during the pilot study, thus extending the scope of the initial work programme for WP6.

First, both WP1 and WP4 underlined the necessity to solve the problem of different metadata formats between institutes and survey programmes. Even if the DDI standard is used, in practice a technical import routine will be needed that supports various metadata formats.

Second, the quality of annotations can be improved by addressing the issue of non-substantive heterogeneity of concept terms observed in WP4. To limit non-substantive heterogeneity, the use of artificial intelligence for concept recommendation can improve both the quality and efficiency of annotation. To explore the potential of a concept recommender system, we developed and evaluated a method for estimating the similarity of questions using "allenai-specter", a SBERT model trained on scientific documents. The results of these models, which identified semantic similarities, were then re-evaluated on a 5-point Likert scale by 3 researchers trained in the social sciences and experienced in survey research. The three annotators had a reliability measured by Krippendorff's α of .82. The Pearson's correlation coefficient between the three annotators is r(1) = .82 with p < .001 in all cases. Thus, we can confirm that the suggestions based on semantic similarities between measures provide a sound basis for recommending concepts. We concluded that a concept recommender system will improve efficiency of annotation. The design of an automated recommender system will require a training corpus of high-quality annotations. However, the design should take into account that automated recommendations carry the potential caveat of obscuring valuable diversity of concept descriptors if users avoid the effort of adding appropriate concepts and instead select only the recommended terms.

Thirdly, at this stage the AnnoTool did not allow annotators to specify relationships between concepts (e.g., "openness" as a sub-dimension of "BIG5"). While these structures partly emerged automatically in the process, a future phase could allow annotators to assign relationships between concepts.

Fourthly, there are use cases where institutions want to register their concepts in bulk, without first aligning them with the existing corpus. They may be interested in linking these new concepts to the existing ones (see the previous paragraph).

Based on the findings from our pilot study, we adapted our process model for a concept registry integrated into a broader LORD infrastructure. Figure 2 summarises the current state of discussion within the project group.

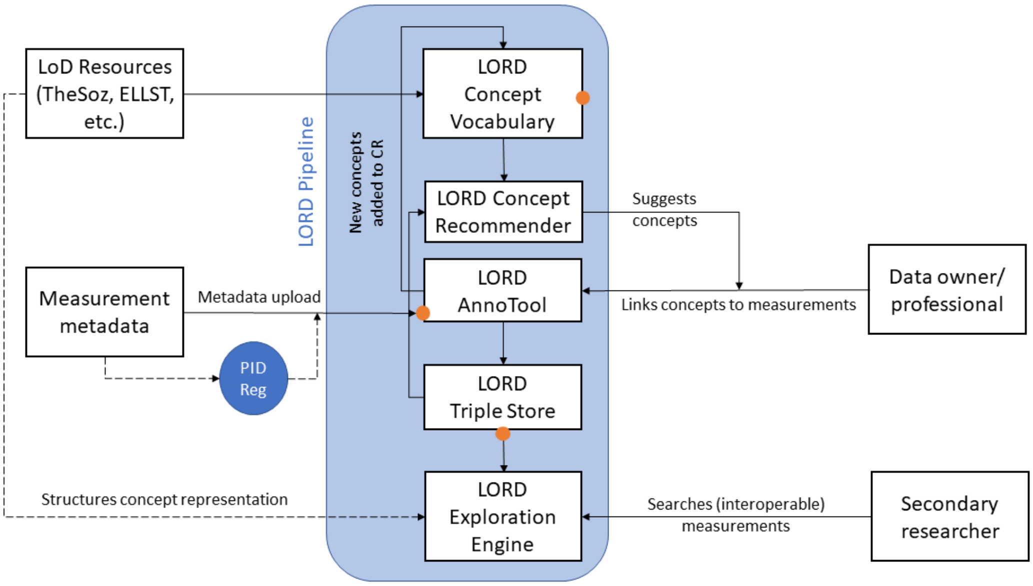

Figure 2: Overview of the LORD infrastructure to be developed in the follow-up project. The filled orange circles are interfaces to access the databases.

The figure shows the components that have been prototyped for the pilot study: (1) the concept vocabulary, (2) the annotation tool, and (3) the triple store. Existing thesauruses and vocabularies will be integrated into the concept vocabulary, as it has been explored with TheSoz[9] for the pilot study.

The first new component is the LORD Concept Recommender, to increase the efficiency of data indexing and limit the heterogeneity of concept terms by suggesting relevant concepts at the time of annotation. The second new component is the LORD Exploration Engine. This component will import the metadata from the LORD Triple Store into a graph-based search index for visually assisted data discovery (see follow-up application). To increase the practical value of the infrastructure, we propose to integrate the PID services for variables developed for the KonsortSWD[10] project into the LORDpipeline, so that persistent identifiers for the measurements can be obtained when the metadata is uploaded.

## Conclusion / Summary

As stated above, the aim of the pilot study was to evaluate whether a user-driven concept registry can be implemented in real-world workflows. Our conclusion is that it is possible if we develop support for semi-automatic indexing, incorporating where appropriate recent advances from Large Language Models (LLMs). Another lesson from the project was that we

[9] https://lod.gesis.org/thesoz/de/, last access on November 16 2023.

[10] https://www.konsortswd.de/angebote/datenzentren/serviceangebote/persistent-identifiers/, last access on November 16 2023.

need to address different use cases for the infrastructure. Some data providers might wish to use all components as a tool for granular metadata management and exploration, others might wish to customise individual components of the infrastructure for their purpose (e.g., only the concept vocabulary).

The results of our project will not only shape the follow-up proposal to be submitted for review soon. The feedback from our presentations at international conferences has underlined the relevance of extending social science terminology services for semantic indexing of research data. The growing number of available research data accompanied by the ongoing professionalisation of data management (especially within large-scale collaborative data collection programmes) and the creation of international search engines for research data are raising awareness of the benefits of controlled vocabularies for data indexing. As always, the main challenges are to design an efficient workflow with a high degree of automation and to secure long-term funding for a product that is a public good for data infrastructures.

The technical components have been prototyped for the purposes of the pilot study. They are not currently available for reuse by other researchers. Full technical development for wider use was beyond the scope of the project. Only the LORD triple store is accessible via the SPARQL endpoint. Our aim is to provide an open access infrastructure if the follow-up application is funded. We will then make all components available for reuse and actively engage with national and international partners to create a shared vision for the future of social science data indexing.

## 4 Publicly Accessible Project Results

### 4.1 Publications with scientific quality assurance

### 4.2 Other publications and published results

- Siegers et al. 2023. "Linked Open Research Data for Social Science – a concept registry for granular data documentation." IASSIST Conference 2023, virtual, 2023-05-31. DOI: 10.5281/zenodo.8090332.
- Nebelin et al. 2023. "Linked Open Research Data for Social Science – a concept registry for granular data documentation." European Survey Research Association (ESRA) 2023 Conference, Milan, 2023-07-18. DOI: 10.5281/zenodo.8232698.
- Siegers et al. 2023. "Linked Open Research Data for Social Science – a concept registry for granular data documentation." 1st Conference on Research Data Infrastructure (CoRDI), Karlsruhe, 2023-09-12. DOI: 10.52825/cordi.v1i.300.
- Siegers et al. 2023. "Linked Open Research Data for Social Science - a concept registry for granular data documentation." ODISSEI Conference 2023, Utrecht, 2023-11-02. DOI: 10.5281/zenodo.10105137.