

# CARDINAL: Contextualized Adaptive Research Data Description INterface Applying LinkedData

André Langer<sup>1</sup>[0000-0001-7073-5377], Christoph Göpfert<sup>1</sup>[0000-0001-6659-8947],  
and Martin Gaedke<sup>1</sup>[0000-0002-6729-2912]

Chemnitz University of Technology, Chemnitz, Germany  
{andre.langer,christoph.goepfert,martin.gaedke}@informatik.tu-chemnitz.de

**Abstract.** To increase the findability of published research data, meta-data has to be provided to describe all relevant characteristics of the contained data. However, common user interfaces for describing research artifacts typically focus on static free-text input elements which constitutes an obstacle for interdisciplinary, unambiguous, fine-grained data descriptions. Reusing already existing domain-specific metadata models based on semantic ontologies are a more promising approach, but the careful selection and presentation of meaningful properties is not trivial. In this paper, we present the CARDINAL approach, which generates and presents an adaptive user input interface to the user that takes the research context of the original digital file into consideration and allows the structured input of knowledge-domain specific descriptive metadata based on existing ontologies. We show in a proof-of-concept the feasibility of such a contextualized web form for research metadata and discuss challenges in the selection process for relevant ontologies and properties.

**Keywords:** Adaptive User Interface, Contextualization, Linked Data, Research Data Management, Data Publishing, Ontologies

## 1 Introduction

In the context of OpenScience, researchers are encouraged to publish their research data (also known as research datasets) in common data repositories so that others can find and reuse them. The term research data refers to any “data being a (descriptive) part or the result of a research process”. Any kind of research literature is usually excluded when using the term research data, e.g., research articles or papers [5, 12].

This research data publishing process shall increasingly be in compliance with the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data management [18]. In the last few years, the FAIR principles have been widely endorsed, such as on national level by the German Research Foundation (DFG) in context of the National Research Data Infrastructure (NFDI), or on global level by the Research Data Alliance (RDA).

As digital research data is normally not self-descriptive, a user has to add additional metadata during the research data publishing process to explicitly

describe all relevant aspects of the contained data to make it findable by search crawlers and other applications. According to the FAIR principles, provided metadata “should be easy to find for both humans and computers.”

Nowadays, data repositories primarily focus on administrative, citation, technical and some basic descriptive metadata [10]. Information on particular data characteristics are either collected not at all, in an unstructured way as floating text or only in domain-specific data repositories. This makes it difficult to simplify the discoverability of relevant datasets for researchers from different knowledge disciplines and results in the current situation, that dedicated scientific search catalogs are relying on keyword-based or fuzzy-logic based full-text search operations in this metadata. And their faceted search possibilities are limited to basic entities such as the *knowledge discipline*, *resource type* or *data license* and certain *provenance* constraints, which is directly in conflict with the FAIR principles to provide rich metadata (F1) in standardized vocabularies (I2) with accurate and relevant attributes (R1).

This is astounding as scientific communities have already yielded domain-specific high-quality, well-structured, controlled vocabularies that contain relevant properties. However, traditional approaches of using static user input interfaces do not take these domain-specific metadata models into account, as static forms are always structured the same way in terms of input elements, neglecting the context of the research artifact being described. A semantic technology-based approach, which focuses on an established metadata representation format and additionally incorporates other relevant vocabularies in such a metadata description, would be a means to improve the interdisciplinary publishing and discovery process.

Within the collaborative research center *Hybrid Societies*<sup>1</sup>, we investigated the realization of an adaptive user input interface for collecting structured, descriptive, detailed research metadata as part of the *PIROL* PhD project [9] and provide the following three contributions:

1. We present CARDINAL to demonstrate an approach on how to select and adaptively incorporate domain-specific properties into a web form.
2. We formalize and discuss the contextualization of the research metadata collection in user input interfaces
3. We show in an online study experiment the acceptance of the approach and the acquisition of additional domain-specific, descriptive metadata.

The results will contribute to the purpose of improving the interdisciplinary findability of published research data.

The rest of the paper is structured in the following way: In section 2, we provide a conceptual problem and requirement analysis based on a comprehensive usage scenario. In section 3, we describe a concept to identify and present relevant properties to a user for metadata input for describing research data. The realization of this process is then shown in section 4 and evaluated in section 5 wrt. acceptance and metadata quality. Section 6 compares our approach with other related literature, and section 7 finally summarizes our results and provides an outlook to future work.

---

<sup>1</sup> <https://hybrid-societies.org/>

## 2 Problem Analysis

The findability and reusability of research data is interrelated as the relevance of search results depends on its suitability for a new application scenario and the possibility to limit a search to particular characteristics of a dataset. By using general-purpose search applications for research data repositories such as the *OpenAIRE search*<sup>2</sup>, *EOSC EUDAT B2FIND*<sup>3</sup> or the *Google Dataset search*<sup>4</sup>, users are already accustomed to a keyword-based input with some basic filter possibilities, where they have to review results on the search pages individually and carefully in order to actually find existing, relevant research data that can be reused or repurposed for their own work beside irrelevant search results. Additional available research datasets might be even existing that will not show up in such a result list as there is a mismatch between the terms used in the meta description of the published research data and the keywords that were entered by a user in a search interface.

Search services for scientific data typically still focus on keyword-based search methods<sup>5</sup>, and filter possibilities are commonly limited to general entities. Instead, it would be a benefit, if a user can make use of more particular filters for characteristics of research data that the user is looking for. This would require better structured metadata that supports concept-based search approaches, but relevant characteristics vary greatly between different knowledge disciplines and a user might not be willing or able to describe all possibly eligible aspects in a research data meta description. A semantic vocabulary-based approach is promising to improve this situation, especially because a large set of domain-specific controlled vocabularies already exists<sup>6</sup>. Research data repositories have started to add support for additional domain-specific structured metadata descriptions, however, the user interface experience is still weak and requires expert knowledge and manual completion as shown in fig. 1, thus, its usage is limited in a broader scope.

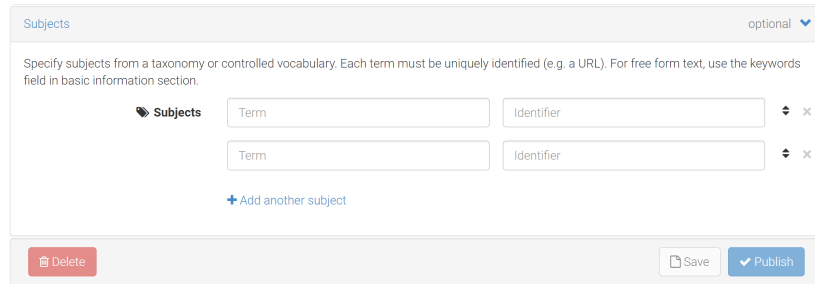


Fig. 1: Current Situation: Provision of domain-specific characteristics, zenodo.org

<sup>2</sup> <https://explore.openaire.eu/search/find>

<sup>3</sup> <http://b2find.eudat.eu/dataset>

<sup>4</sup> <https://datasetsearch.research.google.com/>

<sup>5</sup> <https://www.eosc-hub.eu/services/B2FIND>

<sup>6</sup> <https://lov.linkeddata.es/dataset/lov/vocabs>

In the following, we will concentrate on the metadata acquisition step in a research data publishing process. Therefore, we will describe a fictitious scenario to illustrate an adaptive approach, how a user can be encouraged to provide more specific metadata while maintaining or even improving the user interface experience. *John Doe* is a political scientist and conducts research on electoral behavior. Recently, *John* and his colleagues conducted a randomized survey in which they asked 50 people who they would vote for if they had to choose right now. The answers of the survey participants were compiled in a spreadsheet, which shall now be published to a broader scientific community.

However, depending on the research area which the dataset relates to, further complementary information should be provided. This makes the data easier to interpret and facilitates reuse. In various research areas, there already exist semantic knowledge models about domain-specific concepts in form of ontologies. These can be reused to describe research datasets in more detail. However, concepts which are relevant for research data description need to first be identified. In the provided scenario, an ontology that models characteristics for survey data could be relevant to *John*, such as the *DDI-RDF Discovery (DISCO) Vocabulary*<sup>7</sup>.

There are three user roles to consider for this scenario: The user who publishes research data together with additional descriptive metadata, users that search for existing datasets in the future, and domain experts that provide a domain-specific ontology and the knowledge which concepts are relevant to describe.

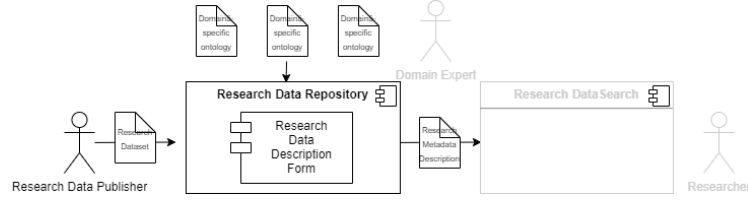


Fig. 2: Conceptual view on the problem scenario

Based on that scenario, we investigate how to design an adaptive submission form for describing a research dataset. Therefore, we identified the following five objectives which also consider the criteria introduced by Paulheim and Probst [13]:

- OBJ1 **Metadata acquisition:** In the user input interface, it shall be possible to provide basic administrative and citation metadata beside additional information based on existing domain-specific ontologies which is relevant to particular research data and might be of interest for other users in the future.
- OBJ2 **Adaptivity:** The form shall be adaptive in the sense that its structure and its components, i.e. input fields, labels, etc., adapt to the context of the research data.
- OBJ3 **Research characteristics:** The form shall reuse existing standardized recommendations for properties and concepts to describe research characteristics.
- OBJ4 **Usability:** The form shall hide technical data details from the user and not create an unsatisfying user interface experience causing additional user effort.
- OBJ5 **Metadata output:** The resulting research data metadata description shall be stored persistently in a machine-readable format, so that it can be easily provided and used in consecutive tool chains, e.g., by corresponding search services.

<sup>7</sup> <https://rdf-vocabulary.ddialliance.org/discovery.html>

### 3 The CARDINAL Approach

The following design is based on three assumptions:

1. A user intends to publish research data, possessing certain attributes that can be related to at least one knowledge domain
2. Standardized vocabularies / ontologies already exist and are available in this knowledge domain in a structured (OWL) description that reflect relevant concepts to describe research in this discipline
3. A subset of these properties is relevant for other users to find and reuse this research data

Input forms are an established means to collect metadata in a manual user input activity. In contrast to static input forms, which are assembled by a developer and commonly present the same input controls to all users, we are heading for an adaptive approach, which will add additional input elements depending on the nature of the resource to describe. Therefore, additional knowledge has to be provided to the application in a first step that can be used to contextualize the further form handling. In a consecutive form building process, relevant ontologies and properties have to be (semi-)automatically selected in order to generate an input form in which a user can then provide and store metadata in a final process step, as shown in fig. 3.

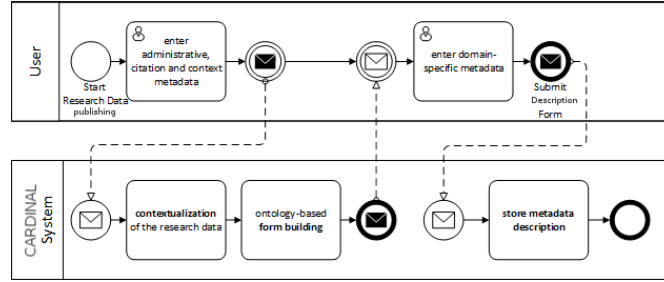


Fig. 3: BPMN process for an adaptive research metadata input form

#### 3.1 Contextualization

In order to tailor the adaptive form to the useful description of particular research data, the context of a published research artifact must be understood. This step is necessary to decide on which domain-specific information is relevant for this research data. We refer to the process of finding a suitable context for an artifact as contextualization.

Referring to context-aware software [16], this can be done based on information explicitly provided by the user, implicitly by processing the provided research data with knowledge extractors if a file artifact is directly provided, implicitly by reusing externally available metadata background information to the dataset if a persistent identifier is already provided, or a combination or variation of the mentioned approaches. For the sake of simplicity, we focus on the first option and reuse meta information that a user might provide anyway when publishing research data (context-triggering actions) independent of a materialized file artifact or identifier.

Contextualization can be related to the classification, characteristics and usage of the investigated object and its origin as well as to spatial or temporal constraints. Attributes that might be used to describe the context of the research data include, for example, the research area(s) which the data can be assigned to, the resource type, or information about the research or application environment in which the data was generated.

Apparently, there exists also a trade-off between the amount of requested metadata for contextualization purposes, the appropriateness of the adaptivity behavior and the effort of the user and the application to achieve the activity result, so these contextualization attributes have to be considered carefully by the application developer.

$$Context(dataset) = \{c_i \mid i \in 0, 1, \dots, c_i \text{ is attribute of dataset with } key(c_i) \text{ and } value(c_i)\} \quad (1)$$

The provided contextual information can then be used to select the most suitable ontologies for a given research data artifact that contain additional characteristics in the meaning of classes and properties that are worth to describe. This decision-making problem can be tackled by strategies such as using a rule-based approach or a decision tree. It is thereby advantageous to limit the value ranges for the contextualization attributes.

### 3.2 Ontology selection

Based on the specified contextual information, relevant ontologies have to be selected which shall be incorporated adaptively into the input form to describe particular characteristics of the research data from the user. As a starting point for finding reusable ontologies, ontology catalogs can be used to search for publicly available ontologies.

In our approach, we suggest a weighted sum model as a simple multi-criteria decision making method for selecting relevant ontologies out of a list of classified available ontologies, shown in eq. (2).

$$\begin{aligned} score(ontology, dataset) &= \sum_{c_i \in Context(dataset)} \omega_{ontology}(c_i) \cdot isMatch(value_{ontology}(c_i), value_{dataset}(c_i)) \\ \text{with } isMatch(value_{ontology}(c_i), value_{dataset}(c_i)) &= \begin{cases} 1 & value_{ontology}(c_i) = value_{dataset}(c_i) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (2)$$

Based on that, an ontology selection component can return all ontologies related to the provided context whose score value exceed a predefined threshold. If no score value exceeds the threshold, an empty list will be returned, thus, the further input form will not be adaptive and not offer additional input fields.

### 3.3 Form Generation

The previously provided ontology selection is used to build the adaptive section of the input form, which we will refer to as ontology-based form building.

In an ideal case, the structured OWL representation of an ontology can automatically be processed to generate an input interface for its provided classes and properties (Code generation through model transformation [8]). However, in practice it turns out, that a direct reuse of ontology representations is not feasible and that an additional presentation specification is needed.

Domain-specific ontologies are usually developed with focus on modeling knowledge about certain concepts, not necessarily with focus on data acquisition. The structure and content of the ontologies can vary greatly. Furthermore, ontologies might contain classes and properties irrelevant for describing research data. However, it is difficult to automate the process to decide which classes and properties are relevant. Instead, this decision should ideally be made in consultation with an expert of the respective domain. Additionally, it might be necessary to describe further layout, order, nesting and repetition possibilities for a certain property. Identifying labels and description texts for input elements is a simple task as long as this information is provided in an ontology. Identifying relevant concepts in an ontology might be done with automated means and additional expert feedback, but a challenging part is to create a reusable layout description of semantic content.

As existing approaches are not applicable in this scenario, we rely on a separately introduced representation of ontologies for presentation and reusability purposes, called *OnForm*, which is described in more detail in [7]. *OnForm* specifications for an ontology are also represented in an RDF format and can be read and interpreted by an *OnForm* generation component. For details on the detailed form generation process, we refer to the corresponding publication.

### 3.4 Metadata Acquisition

After generating a form user interface based on an *OnForm* description, a user can then enter context-specific information into the form fields that will be stored as metadata to describe this research data.

The input elements can make use of additional information provided by the respective ontology, such as an *rdf:type* or *rdfs:range* constraint, in order to render input elements with assistance, entity lookup and auto-completion functionality to increase the user interface experience, hide on a technical level semantic persistent identifiers from the user and nevertheless collect structured, unambiguous information [11].

### 3.5 Metadata Persistence

Following the completion of an adaptive research data submission form, a serialized metadata description has to be created based on the provided form input data. Using established semantic technologies, this process can then be done in a straight-forward fashion as the user interface itself is already based on RDF classes and properties with corresponding identifiers. The provided values by the user will be taken by a persistency component and stored in a common RDF serialization format such as RDF/XML, JSON-LD or Turtle.

## 4 Prototypical Design

Based on the concept presented in the previous section and our sample application scenario from section 2, we designed a prototypical CARDINAL application for Creating an Adaptive Research Data Description Interface that applies Linked Data properties and concepts based on existing ontologies..

It realizes an adaptive web form, that is divided in a straight-forward fashion into three sections as depicted in fig. 4. Each serves a distinct purpose. The first section requires users to provide general literal administrative and citational metadata. The second section requests additional common meta information that constitutes the basis for contextualizing the research data-specific further part of the web form. The third section is then built adaptively depending on the selected *OnForm* ontology.

In this simple form, the user does not have to provide the research data itself or any reference to it, as we do not focus on automated classification and knowledge extraction methods in this paper.

In order to select appropriate attributes for the contextualization section, we carefully reviewed existing user interfaces of established research data repository providers, namely the research data submission forms from *Zenodo*, *ResearchGate*, *Mendeley Data*, *Dataverse* and *B2SHARE*. In all of these application, a dedicated input field for *keywords* and the *file type* is already established and users are used to provide this additional information. Additionally, we add the *research area* as another contextual attribute as this information might be directly related to existing vocabularies established by dedicated communities.

In order to limit the value range of these contextual attributes to mappable characteristics of existing ontologies, we rely on existing classifications schemes for these attribute values. As a basis for suggesting research areas, a comparison of existing librarian classification systems focusing on scientific publications was done. As a result, the German/Dutch Basisklassifikation (BK) was used, which offered a number of 48 main classes and was already available in a structured RDF description, which made it simple to integrate the provided research area resource URI into a meta description. For the file type, our review also resulted in a set of typical data types for our demonstrator, containing *Audio*, *Code/Software*, *Document*, *Image*, *Model*, *Tabular Data*, *Text*, and *Video* similar to the recommended list for *DCMITypes*. We excluded common generic data types such as *Dataset* or *Publication*, as their usage for contextualization was considered limited. The scope of allowed keywords is difficult to limit in practice. As we focus on identifiable concepts with a persistent mappable identifier, we added an auto-suggestion feature to the keyword input element which retrieves keyword suggestions from an appropriate terminology, such as *DBpedia*<sup>8</sup>, *Wikidata*<sup>9</sup> and the *WordNet*<sup>10</sup> dump.

---

<sup>8</sup> <https://dbpedia.org/sparql>

<sup>9</sup> <https://query.wikidata.org/>

<sup>10</sup> <https://wordnet.princeton.edu/>



### 1. Citational Information

Title	<input type="text" value="Survey on Electoral Behavior"/>	?
Creator	<input type="text" value="John Doe"/>	?
Description	<input type="text" value="Results of a phone survey on electoral behavior with 50 partici"/>	?
Year of Publication	<input type="text" value="2020"/>	?

### 2. Context

Research Area	<input type="text" value="Social Sciences"/>	?
Data Type	<input type="text" value="Tabular Data"/>	?
Keywords	<input type="text"/> <p>Please select a suggested keyword. Submitted keywords are displayed below.</p> <div> <span>survey x</span> <span>questionnaire x</span> <span>politics x</span> <span>election x</span> </div>	?

### 3. Survey

Question	<input type="text" value="If there were federal elections next Sunday, which party would"/>	?						
Answer or Variable	<input type="text" value="A party of the German Bundestag."/>	?						
Analysis Unit	<input type="text" value="political party"/>	?						
Period of Time	<table> <tr> <td>Start Date</td> <td><input type="text" value="07 / 06 / 2020"/></td> <td>?</td> </tr> <tr> <td>End Date</td> <td><input type="text" value="07 / 10 / 2020"/></td> <td>?</td> </tr> </table>	Start Date	<input type="text" value="07 / 06 / 2020"/>	?	End Date	<input type="text" value="07 / 10 / 2020"/>	?	
Start Date	<input type="text" value="07 / 06 / 2020"/>	?						
End Date	<input type="text" value="07 / 10 / 2020"/>	?						

Fig. 4: Exemplary user interface for a generated adaptive web form

Available domain-specific ontologies were retrieved from ontology catalogs, such as *Linked Open Vocabularies (LOV)*<sup>11</sup>. The retrievable ontologies are already tagged with basic category labels that were considered for the ontology selection process in the CARDINAL prototype, in such a way, that we curated a list of relevant ontologies and stored for each of these ontologies a context definition, attribute weight and *OnForm description* as a basic application configuration.

We followed our introductory scenario example and added an *OnForm* representation of the *DISCO* ontology<sup>12</sup> together with matching for research data with keywords, such as *Survey* or *Questionnaire* and a data type *tabular data*

<sup>11</sup> <https://lov.linkeddata.es/dataset/lov/vocabs>

<sup>12</sup> <http://purl.org/net/vsr/onf/desc/survey>

or *text*, independent of the research area. Similar rules can, of course, also be defined for other usage scenarios, e.g., for applying a multimedia ontology to describe an *image*, *video* or *3d model*.

After a user fills out and submits the generated form, all form data is stored and provided in Turtle as an RDF serialization format for download. An example is given in fig. 5. We emphasize, that the information in the highlighted section is additionally gathered by the adaptive CARDINAL approach in comparison to traditional research data submission forms.

```

1 @prefix dc: <http://purl.org/dc/elements/1.1/>
2 @prefix dcterms: <http://purl.org/dc/terms/>
3 @prefix dctype: <http://purl.org/dc/dcmitype/>
4 @prefix dbr: <http://dbpedia.org/resource/>
5 @prefix disco: <http://rdf-vocabulary.ddialliance.org/discovery#>
6 @prefix ex: <http://www.example.org/>
7
8 ex:dataset1
9   dc:title "Survey on Electoral Behavior";
10  dc:creator "John Doe";
11  dc:description "Results of a phone survey on electoral behavior with 50 participants";
12  dc:date "2020";
13  dct:DCMIType dctype:Text;
14  dc:subject dbr:Social_Sciences, dbr:Survey_(human_research), dbr:Questionnaire, dbr:Politics, dbr:Election.
15
16  rdf:type disco:Questionnaire ;
17  disco:question [
18    rdf:type disco:Question ;
19    disco:questionText "If there were federal elections next Sunday, which party would you give your vote?";
20    disco:variable [
21      rdf:type disco:Variable;
22      disco:description "A party of the German Bundestag." ;
23      disco:analysisUnit "political party"^^disco:AnalysisUnit.
24    ];
25    disco:temporal [
26      rdf:type disco:PeriodOfTime;
27      ns2:startDate "2020-07-06"^^xsd:date ;
28      ns2:endDate "2020-07-10"^^xsd:date .
29    ]
  ]

```

Fig. 5: Exemplary metadata export result in Turtle

## 5 Evaluation

In order to evaluate our proposed approach, we implemented the designed prototype from section 4 as a proof-of-concept<sup>13</sup> in *Python* based in a straight-forward fashion on *Django*, *Bootstrap* and *rdflib*.

The demonstrator was then used in an unsupervised online study which contained a web-based survey experiment. The survey was realized with *LimeSurvey*<sup>14</sup> and distributed via university mailing lists and the platform *SurveyCircle*<sup>15</sup>. Based on the objectives defined in section 2, the online study had the purpose to analyze the feasibility and acceptance of an adaptive input form approach. We were therefore especially interested in the extent of the acquired metadata and its data quality characteristics, additional user effort to complete the adaptive form section as well as occurring usability issues.

The study was based on a given fictitious initial scenario, similar to section 2. It was provided to all study participants at the beginning of the survey description in German or English.

<sup>13</sup> <http://purl.org/net/vsr/onf/onform>

<sup>14</sup> <https://bildungsportal.sachsen.de/umfragen/limesurvey/index.php/877377>

<sup>15</sup> <https://www.surveycircle.com/>

The participants were randomly divided into two groups. The participants of group A were given a traditional static submission form which did not contain an additional research data context-specific section whereas the participants of group B saw the adaptive submission form with an additional dynamic form section based on their contextual selection. After participants completed the input procedure, they were asked to fill out a System Usability Score (SUS) questionnaire to evaluate the usability of the system.

The survey took place without incentives over a period of one month between July 2020 - August 2020 and reached 83 participants with 74 full responses<sup>16</sup>. The majority of our participants assigned their primary field of knowledge to Economics (38), Psychology (15) and Social Science (7), but the target group also contained participants from the field of Engineering & Computer Science (6), Medicine (2), and other disciplines with varying age and experience level. 35 participants were assigned to version A, thus, the static submission form, and 39 participants were assigned to version B, the adaptive submission form.

### 5.1 Acquired metadata

For test group B, the entered general information from the study participants was used to identify a suitable ontology together with a corresponding *OnToForm* description and to display in a separate form section additional input fields to the user. Based on the provided input data, we evaluated the extent, to which the contextualization step worked as expected and if additional metadata was actually entered by the adaptive test group B in comparison to the reference group A.

Identifying a context based on the provided keywords worked very well. Within the experiment, we relied on *WordNet* as a data source and compared the entered keywords with it. A total of 115 keywords were provided. 86 out of 115 keywords were selected from suggestions, so that internally the user input could be mapped to a corresponding resource URI successfully. Surprisingly, explicitly providing a data type for the given scenario was unexpectedly challenging for the participants. Although the scenario description stated to publish an *Excel spreadsheet*, only 23 users actually selected the tabular data option, whereas other participants selected Document (6), Text (6) or even Audio (4). Especially in the last case, the form did not adapt as intended to the context of questionnaire data. In the following form section, participants were then able to specify information based on the context-related *DISCO* ontology. Figure 6 shows which of these input fields participants completed or skipped. A question text was provided by 33 participants. All of these participants either correctly reused the question from the task scenario or rephrased it slightly. The field for "Analysis Unit" was skipped the most with only 26 participants specifying a value. This might originate in the ambiguous label provided by the ontology and could be resolved in the future by specifying a better term in the *OnToForm* representation that is more comprehensive for users.

Only two participants decided to skip the second form section entirely.

---

<sup>16</sup> <https://doi.org/10.5281/zenodo.4439700>

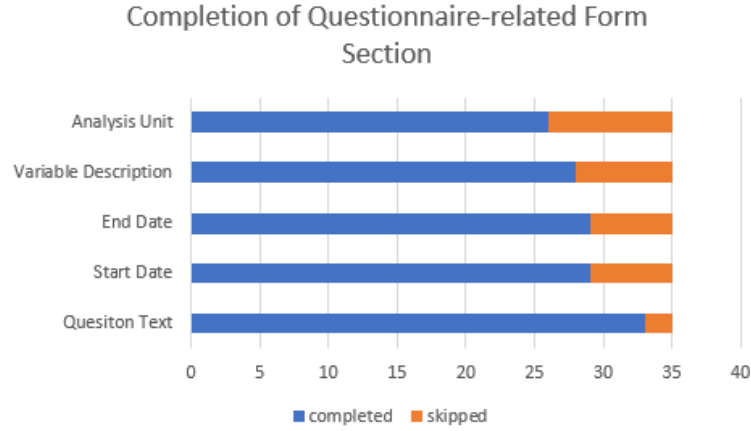


Fig. 6: Number of participants in test group B that provided additional descriptive metadata

## 5.2 User Effort

Additionally, we measured Time to Completion. The adaptive form consisted of 12 fields compared to 4 fields of the static form – therefore, we expected that the completion time would approximately triple as well, which turned out to be correct as shown in table 1.

	Min Time	Max Time	Avg Time
Group A	16s	72s	101s
Group B	58s	1083s	286s

Table 1: Metadata form input completion times for static form (group A) and adaptive form (group B)

Nevertheless, we want to point out its relevance, that the adaptive approach will only display input elements to the user that are worth to consider in comparison to an input form which simply displays all imaginable input elements to a user.

## 5.3 Usability assessment

We used a System Usability Score (SUS) questionnaire as introduced by Brooke to measure usability [2]. The static form thereby received an average SUS score of 72 (standard deviation 15.88) and the adaptive form a score of 58 (standard deviation 16.95). According to the Sauro-Lewis curved grading scale for the SUS, the static form is located in the 60-64 percentile range and the adaptive form is located in the 15-34 percentile range [15]. In terms of scores, the static form achieved a score of C+, far exceeding the adaptive form, which is rated D. This indicates a below-average user-friendliness.

## 5.4 Discussion

Our objective was to develop an adaptive description form to obtain more detailed metadata descriptions of research data in a machine-readable semantic format than currently possible. For this purpose, we were able to reuse classes and properties of an existing ontology (*DISCO*) to annotate research data that contains questionnaire data. The evaluation results show that this objective has been achieved. Out of 35 users, only 2 users did not provide additional metadata in the adaptively generated form section. That means that in 91.4% of all submission procedures (32 out of 35 cases) our approach has led to an improved metadata description of the research data. Our approach requires users to specify additional context metadata to contextualize research data. Nevertheless, usability problems do still exist. Users are able to use the adaptive description form efficiently and effectively, but they are evidently not satisfied with it. According to user feedback, this is mostly due to labels and help texts that are incomprehensible to some. Therefore, focus should be put on improving usability in the future.

## 6 Related Work

The use of domain-specific standards for providing metadata descriptions for research data is highly encouraged by groups such as force11 [18] or the Alliance of German Science Organizations. Multiple standards were already established for certain types of research data across different research areas [3, 19]. Due to the large number of available standards, it is not feasible to create a holistic metadata description form that contains input fields to provide metadata for every single existing standard.

In the field of human-computer interaction, research is conducted on context-aware computing and context-aware software. Schilit et al. [16] use the term in reference to mobile computing by mainly focusing on the physical environment of devices. They state that context-awareness software can be implemented by using simple if-then rules. In contrast, Schmidt et al. [17] argue that the term context refers to more than just the physical environment. They classify contexts into two categories: human factors and physical environment.

There are various suggestions to define the term adaptivity. Preim and Dachselt [14] distinguish three types of adaptivity wrt. the experience of users, perceptual skills and work environments. At present, there are no research data repositories that adapt to the context of a research data artifact. However, a few approaches have been proposed that show how adaptive user interfaces can be designed. Baclawski and Schneider implemented an approach for using ontologies to describe data [1] by annotating their research data with additional metadata. For that, they developed the Open Ontology Repository (OOR) infrastructure to manage the "metadata management life-cycle". However, their system relies heavily on user expertise. OOR provides functionality for users to rank and tag ontologies. Users are required to determine themselves which ontology is most suitable for their research data.

A similar initial situation is described by Cimino and Ayres [4]. For their solution approach, they developed a common repository for the administration of clinical research data. Their solution system is named Biomedical Translational Research Information System (BTRI) and also makes use of ontologies, similar to the approach of Baclawski and Schneider. Besides using ontologies to describe data in more detail, Gonçalves et al. have shown an approach on how ontologies can be used to generate web forms. They introduced an "ontology-based method for web form generation and structured data acquisition" [6]. Their system requires two input files to generate a web form which is used to digitize a standardized questionnaire. Firstly, an XML file is used to configure the form layout, as well as bindings of user interface components to entities within the ontology. Secondly, a form specification is used to define the actual content of the form. Paulheim and Probst created an extensive state of the art survey on ontology-enhanced user interfaces [13]. An ontology-enhanced user interface is defined as "a user interface whose visualization capabilities, interaction possibilities, or development process are enabled or (at least) improved by the employment of one or more ontologies".

Although there are currently no research data repositories that employ adaptive user interfaces as defined in this section, the use of adaptive user interfaces can be advantageous to provide users with means to describe their research data in more detail, specifically by using domain-specific terminology. These user interfaces should adapt to the context of a user's research data artifact. With regard to the definition of adaptivity according to Preim and Dachsel, this corresponds in the broadest sense to the adaptivity type with regard to the physical environment. However, we instead refer to the contextual environment of research data and not to the physical environment as in the original definition. As exemplary case studies of [4, 1, 6] show, emphasis should be placed on data interoperability. For this purpose, ontologies for describing the structure and semantics of data proved to be useful. Furthermore, the extensive state of the art survey of Paulheim and Probst provides many more cases that show how ontologies can be used to enhance user interfaces.

## 7 Conclusion

In this paper, we focused on the description step in research data publishing processes and discussed an adaptive ontology-based form building approach based on existing, domain-specific ontologies in order to provide metadata descriptions with additional context-related structured meta information. By reusing existing ontologies, our approach relies on controlled expert knowledge stating which aspects are important to describe in a specific application domain. By considering general contextual information for the selection of relevant ontologies, we relieve the user from filling out extensive research data submission forms with input elements that are not relevant at all as well as the developer who had to manually craft detailed input forms in the past. The additionally acquired metadata can facilitate the interdisciplinary findability and reuse of existing research data based on Linked Data.

We implemented our suggested CARDINAL approach, demonstrated it as a proof-of-concept and additionally evaluated the solution based on an on-line survey experiment with 83 participants. The results of our user study proved that the prototype could successfully be used for obtaining more detailed metadata descriptions. The results also showed that our prototype fulfills all of the requirements apart from usability weaknesses, where the survey results already disclosed some issues.

As future work, it is necessary to further improve the user interface experience as well as to provide further *OnForm* descriptions for existing domain-specific ontologies. Furthermore, it makes sense to investigate more deeply possibilities to semi-automatically create these *OnForm* representations by applying more sophisticated ontology classification algorithms, property relevance metrics and knowledge extraction methods.

**Acknowledgment.** This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 416228727 – SFB 1410

## Bibliography

1. Baclawski, K., Schneider, T.: The open ontology repository initiative: Requirements and research challenges. In: In Proceedings of Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, ISWC (2009)
2. Brooke, J.: Sus: A quick and dirty usability scale. Usability Eval. Ind. 189 (11 1995)
3. Büttner, S.: Langzeitarchivierung von forschungsdaten. eine bestandsaufnahme. Information - Wissenschaft und Praxis 65(4-5), 299 – 300 (01 Sep 2014)
4. Cimino, J., Ayres, E.: The clinical research data repository of the us national institutes of health. studies in health technology and informatics, 160, 1299-1303. Studies in health technology and informatics 160, 1299-303 (01 2010)
5. Elsevier: Sharing research data (2021), <https://www.elsevier.com/authors/author-resources/research-data>
6. Gonçalves, R., Tu, S., Nyulas, C., Tierney, M., Musen, M.: An ontology-driven tool for structured data acquisition using web forms. Journal of Biomedical Semantics 8 (08 2017)
7. Göpfert, C., Langer, A., Gaedke, M.: Ontoform: Deriving web input forms from ontologies. In: Web Engineering - 21th International Conference, ICWE 2021, Biarritz, France, May 18-21, 2021, Proceedings. Lecture Notes in Computer Science, Springer (2021), currently under Review
8. Hemel, Z., Kats, L.C., Groenewegen, D.M., Visser, E.: Code generation by model transformation: A case study in transformation modularity. Software and Systems Modeling 9(3), 375-402 (2010)
9. Langer, A.: PIROL : Cross-domain Research Data Publishing with Linked Data technologies. In: La Rosa, M., Plebani, P., Reichert, M. (eds.) Proceedings of the Doctoral Consortium Papers Presented at the 31st CAiSE 2019. pp. 43-51. CEUR, Rome (2019)
10. Langer, A., Bilz, E., Gaedke, M.: Analysis of current RDM applications for the interdisciplinary publication of research data. CEUR Workshop Proceedings, vol. 2447. CEUR-WS.org (2019)
11. Langer, A., Göpfert, C., Gaedke, M.: URI-aware user input interfaces for the unobtrusive reference to Linked Data. IADIS International Journal on Computer Science and Information Systems 13(2) (2018)
12. Pampel, H., Vierkant, P., Scholze, F., et al.: Making research data repositories visible: The re3data.org registry. PLOS ONE 8(11), 1-10 (11 2013)
13. Paulheim, H., Probst, F.: Ontology-enhanced user interfaces: A survey. Int. J. Semantic Web Inf. Syst. 6, 36-59 (04 2010)
14. Preim, B., Dachselt, R.: Interaktive Systeme: Band 1 Grundlagen, Graphical User Interfaces, Informationsvisualisierung. eXamen.press, Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg (2010)
15. Sauro, J., Lewis, J.R.: Quantifying the User Experience, Second Edition: Practical Statistics for User Research, vol. 38 (2016)
16. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: 1994 First Workshop on Mobile Computing Systems and Applications. pp. 85-90 (1994)
17. Schmidt, A., Beigl, M., Gellersen, H.W.: There is more to context than location. Computers and Graphics 23(6), 893 – 901 (1999)
18. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The fair guiding principles for scientific data management and stewardship. Scientific Data 3(1), 160018 (Mar 2016)
19. Woolcott, L.: Understanding metadata: What is metadata, and what is it for?., Cataloging & Classification Quarterly 55(7-8), 669-670 (2017)