BIRD: Using Conversational User Interfaces to Provide Relevant Metadata for Interdisciplinary Research Data Publishing

André Langer, Lukas Schmolke and Martin Gaedke

Professorship for Distributed and Self-organizing Systems,
Chemnitz University of Technology, Germany

By the digitization of science, publishing research data to the scientific community is increasingly demanded in order to allow the replay, reuse or repurpose of existing research results. It is often necessary to describe the original research data files to increase its findability according the the FAIR principles for good scientific practice. So far, this is typically done by providing an additional descriptive floating text or by stating specific information in a submission form of a research data repository. Alternatively, structured machine-readable metadata description files can directly be provided. However, the meta description result quality depends on the experience of the user, commonly focuses on general aspects, and tool assistance is often limited. To address these issues, we applied as an alternative a conversational user interface paradigm to the description of research data and present the BIRD prototype (Bot-based Interface for Research Descriptions). It realizes a chat dialog which will assist scientists in providing an appropriate, structured metadata description based on the OpenAIRE Guidelines for Data Archives, independent of a particular technical repository platform. The tool is offered as a demonstrator for public access.

1 Introduction

When publishing scientific artifacts, such as recorded files from an experiment, generated files from a software, or developed application components, researchers are encouraged to provide additional structured meta information about certain characteristics of these scientific datasets, as these are normally not self-descriptive. For that purpose, several proposed metadata standards and schemas already exist [1]. Such a metadata description nowadays commonly comprises a title, some information about the author and institution, some other administrative or citation metadata, some simple and maybe ambiguous keywords and an unstructured free-text description of the main content. However, especially for early-career researchers, it is an obstacle to start with research data publishing

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029704 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13745 (CC BY-SA 4.0)

(RDP) because they might not be aware of relevant existing standards, are bored to fill out extensive, static, text input-oriented submission forms in well-established research data repository applications, or see it as a time-consuming activity without support or interaction.

In 2020, we conducted a survey among early 24 career researchers of varying discipline, experience level and demographic characteristics at Chennitz University of Technology, which focused on their research data publishing and metadata description behavior [2]. As a result, it is shown in Fig. 1, that a vast amount of the participants has never published research data with an appropriate description so far and that knowledge about existing standards on schemas and vocabularies is limited.

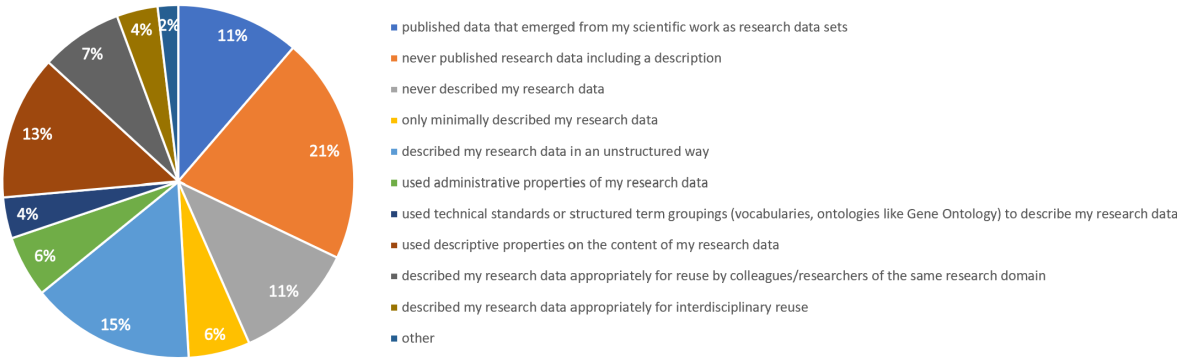

Figure 1: Preliminary survey among 24 early-career researchers on research data publishing.

Furthermore, the findability and filter-ability for particular research data requires highly structured, unambiguous, fine-grained metadata, which is often not provided in floating text data descriptions or free text form fields, especially for interdisciplinary purposes.

To address this issue, Chatbot-like user interfaces are a promising approach that were already successfully applied in other knowledge domains to request structured information from a user and guide the user through a set of relevant questions in an adaptive fashion [3]. In the particular domain of scientific metadata management, the number of existing approaches is still limited.

We investigate the feasibility and challenges of such a conversational UI-based approach to build the prototype of a dialog system for research data descriptions based on the Rasa framework and the OpenAIRE Guidelines for Data Archives which will generate a semantically enriched XML file result. This export file can then be used as a structured data source in a consecutive application or tool chain, or it can simply be published as microdata together with the corresponding dataset on web platforms, in order to improve the structured description and discoverability of the shared research data according to the FAIR principles. This research is part of the PIROL PhD project on Publishing Interdisciplinary Research Over Linked data [4] and carried out in the context of the highly interdisciplinary collaborative research center SFB1410 Hybrid Societies.

The rest of the paper is structured in the following way: In section Related Work, we discuss existing metadata standards for research data meta descriptions, user interface

approaches on how to provide them and related work on the application of conversational user interfaces to that scenario. Section BIRD Concept states objectives and the general concept to realize such a description tool. The implementation of a prototypical demonstrator is discussed in section Realization. Section Conclusion summarizes our results and provides an outlook to future activities.

2 Related Work

Providing an additional metadata description for published research data can be done in various ways, as stated in [5], encompassing also traditional approaches based on simple, rarely structured readme plain text files. Encouraged by the FAIR Principles from [6], vocabularies can alternatively be used to provide such a metadata description in a structured way, such as DCAT-AP [7], the OpenAIRE Guidelines [8] for Data Archives based on DataCite [9] or schema.org/Dataset [10]. Beside these general-purpose controlled vocabularies, more than 1.500 other ontologies are listed in 2021 in [1].

Tools already exist that support users in the provision of required meta information based on these vocabularies, which are typically form-based submission interfaces in research data management applications, wizard-based approaches, such as [11], or markup generators [12, 13].

Chatbots are an alternative, already well-understood paradigm to interact with the user. They were already successfully applied in several knowledge domains [14] and studies also show benefits in the application of a dialog-based system in the acquisition of scientific data [15, 3]. Intelligent Conversational User Interfaces as an advancement are expected to be one of the most rapid growing markets within the next years [16]. To the best of our knowledge, no contribution exists so far that demonstrated the feasibility of an applied chatbot interface for generating platform-independent research metadata descriptions. Thus, we will do that in the following.

3 BIRD Concept

Based on the hypothesis, that especially researcher groups with limited previous experience in publishing research data can benefit from a chatbot-based, ontology-considering dialog system, we formulate the following five objectives:

OBJ1 **Dialog - based user interaction**

OBJ2 **Collection of descriptive research metadata**

OBJ3 **Adaptive conversation progress**

OBJ4 **Possibility for additions and changes**

OBJ5 **Structured, platform-independent result export**

To address OBJ1, we focus in a straight-forward fashion on a rule and story-based chatbot interface to facilitate the creation of FAIR research data meta descriptions with an emphasis on descriptive metadata decoupled from a particular application.

We exemplarily base it for OBJ2 on the DataCite class and property definitions and OpenAIRE guidelines to realize the main interaction path ("Happy Path").

A basic dynamic interaction (OBJ3) with the user is realized with alternative select options for particular questions in the dialog and with the possibility to rewind and correct the last action (OBJ4), or by simply requesting additional explanation.

A structured, schema-based XML metadata export functionality is provided for OBJ5.

4 Realization

We implemented the concept as a Proof-of-Concept (PoC) and published the BIRD prototype for public access[1]

The implementation of an intent-based solution for general metadata was straight-forwardly realized with slots. In the backend, we have chosen the OpenSource Python-based Rasa[2] Natural Language Processing (NLP) and Understanding (NLU) framework for realizing the logic of the chatbot based on a dedicated action server. The frontend user interface was realized based on React/SocketIO/NodeJS and a rasa-webchat component[3] as depicted in figure 2.

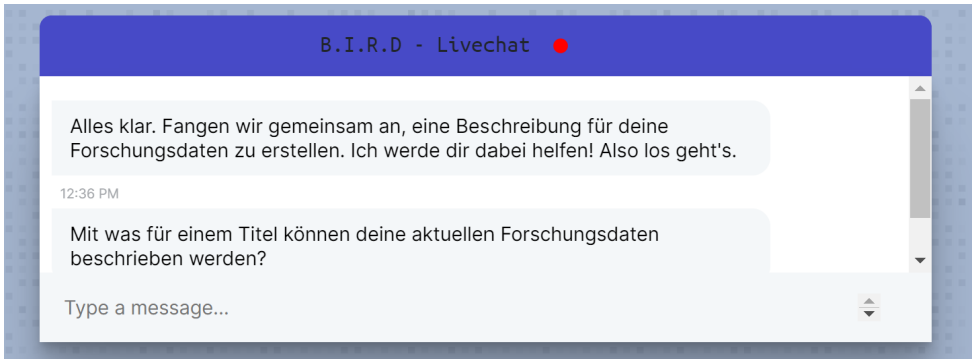

Figure 2: Basic BIRD UI as React Rasa-Webchat component (with German content).

Using intents in this scenario turned out to be challenging, as the chatbot asked the question in an active fashion, and the provided user answer had to be carefully interpreted and filtered. NLP processing performance was OS-dependent and the framework documentation had pitfalls. Handling robustness and a posteriori corrections was challenging and only partially practicable.

[1] https://www.pirol-data.de/bird.
[2] https://rasa.com/
[3] https://github.com/botfront/rasa-webchat

5 Conclusion

To foster the interdisciplinary publication and discovery of research data, tools with a better user interface experience have to be developed that allow a natural and effective provision of structured, relevant meta information with limited prior knowledge.

The BIRD prototype shifts away from traditional forms and allows a dialog-based textural or even lingual metadata collection.

After deployment, it is going to be assessed in a real-world usage scenario. The main focus is then on incorporating taxonomical persistent semantic concept identifiers for common research characteristics and to improve the adaptive dialog behavior. Alternatively, it is worth to investigate hybrid approaches that combine a conversational user interface with a web form.

Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) Project ID 416228727 SFB 1410.

ORCID IDs

- André Langer https://orcid.org/0000-0001-7073-5377
- Martin Gaedke https://orcid.org/0000-0002-6729-2912

Bibliography

[1] FAIRsharing Standards Overview page. `https://fairsharing.org/standards/`, Accessed: 2021-04-22.

[2] Matthias Tietz, André Langer, and Martin Gaedke. Survey results on the interdisciplinary description of research data. `https://purl.org/net/vsr/storch/survey`, Accessed: 2021-04-22.

[3] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. `https://doi.org/10.1145/3290605.3300316`

[4] André Langer. PIROL : Cross-domain Research Data Publishing with Linked Data technologies. In Marcello La Rosa, Pierluigi Plebani, and Manfred Reichert, editors, *Proceedings of the Doctoral Consortium Papers Presented at the 31st CAiSE 2019*, pages 43–51, Rome, 2019. CEUR.

[5] Dong Joon Lee and Besiki Stvilia. Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLOS ONE*, 12(3):1–44, 2017. https://doi.org/10.1371/journal.pone.0173987.

[6] Mark D. Wilkinson, Michel Dumontier, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. https://doi.org/10.1038/sdata.2016.18.

[7] Bert Van Nuffelen. DCAT Application Profile for data portals in Europe Version 2.0.1. 2020. URL: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

[8] Pedro Príncipe, Najla Rettberg, Eloy Rodrigues, Mikael K. Elbæk, Jochen Schirrwagen, Nikos Houssos, Lars Holm Nielsen, and Brigitte Jörg. Openaire guidelines: Supporting interoperability for literature repositories, data archives and cris. *Procedia Computer Science*, 33:92–94, 2014. 12th International Conference on Current Research Information Systems, CRIS 2014. URL: https://www.sciencedirect.com/science/article/pii/S1877050914008059, https://doi.org/https://doi.org/10.1016/j.procs.2014.06.015.

[9] Noémie Ammann, Lars Holm Nielsen, Sebastian Peters, and Madeleine de Smaele. Datacite metadata schema for the publication and citation of research data. 2011.

[10] Data and Datasets - schema.org. https://schema.org/docs/data-and-datasets.html, Accessed: 2021-04-22.

[11] KNB. Morpho, data management for earth, environmental and ecological scientists. https://knb.ecoinformatics.org/tools/morpho, Accessed: 2021-04-23.

[12] DataCite Metadata Generator - Kernel 4.3. https://dhvlab.gwi.uni-muenchen.de/datacite-generator/, Accessed: 2021-04-22.

[13] NustartSolutions. Awesome Step-by-Step JSON-LD Schema Generator Tool (2019). https://nustart.solutions/tools/json-ld-schema-generator-tool/, Accessed: 2021-04-23.

[14] Stefano Valtolina, Serena Di Gaetano, and Pietro Diliberto. Chatbots and Conversational Interfaces: Three Domains of Use. Technical report, 2018. URL: http://ceur-ws.org.

[15] Irene Celino and Gloria Re Calegari. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies*, 139:102410, 2020. URL: https:

//www.sciencedirect.com/science/article/pii/S107158192030015X, https://doi.org/https://doi.org/10.1016/j.ijhcs.2020.102410.

[16] Mordor Intelligence. Chatbot Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026). https://www.researchandmarkets.com/reports/4622740/chatbot-market-growth-trends-covid-19-impact, Accessed: 2021-04-22.