PIROL: Cross-domain Research Data Publishing with Linked Data technologies

André Langer[0000−0001−7073−5377]⋆

Chemnitz University of Technology, Germany
andre.langer@informatik.tu-chemnitz.de

Abstract. Effective research data management for traceability, preservation and reuse is an important part of good scientific practice and is already under discussion over a long period of time. However, the digital transformation in science also led to new challenges for researchers on how to describe, publish and share their research data. This includes the interdisciplinary annotation and discovery of research data, data privacy issues in exposure of data with trends to decentralized platforms as well as sophisticated automatisms to ensure data quality and compliance aspects. Only limited tool support exists for these processes so far.
The following research project will use Linked Data principles to improve the current situation in this problem domain. It will first focus on components and services, that assist researchers in the annotation process of their research data. Next, it will investigate how this research data can be stored and discovered in decentralized, multi-user scenarios to allow data reuse under respect of data privacy concerns. In a third step, meta data descriptions will be used to apply automated data conformance and quality assessment operations on scientific data.

Keywords: Data Annotation, Data Publishing, Decentralization, Data Quality, Linked Data, SoLiD, Open Research

1 Introduction

The term digital research data (sometimes also referred to as scientific data or scholarly data) covers in principle any kind of digital artifact that is associated with scientific research [9].

Research data is an essential artifact of scientific work: It leads to insights, makes research reproducible and validates findings. Therefore, it is inseparable connected to the results of research and also forms the base for future activities. The digital transformation of science has risen new opportunities and issues for all involved stakeholders on national and international level on how to annotate, publish, archive, find and reuse digital research data. Researchers are confronted with new challenges when aiming to publish and share or reuse research data digitally. This applies especially for multifaceted research data where researchers from many cross-domain knowledge domains work together in a collaborative fashion.

⋆ This research is developed under the supervision of Prof. Dr.-Ing. Martin Gaedke.

The process of research data management is commonly described in a data life-cycle. Griffin et al. present a researcher-focused approach and describe five phases for the complex process of research data management, which affect each other. Several modifications and interpretations of this life-cycle can be found. Differences primarily focus on the introduction of a separate planning phase for research data, between the discovery of existing data and the generation of new research data as depicted in fig. 1. Within this data life-cycle, the intended PIROL research project will focus on the publication process (6) of research data. This will obviously also be influenced by the storage of data and have an impact on the discovery.

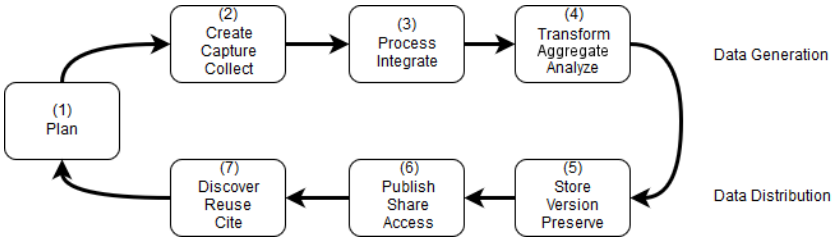

Fig. 1. Research Data Management Life-cycle

Common online publication channels reach from direct data exchange, traditional self-hosted webserver downloads, up to larger centralized platform infrastructures and public/private cloud-based solutions with corresponding application providers. Projects such as Google's Dataset Search, OpenAIRE or Re3Data.org show, that there are already several data publication platforms that provide datasets with basic meta descriptions. Many of them also consider the usage of semantic meta data descriptions, persistent identifiers and concept URIs. Nevertheless, ongoing activities in projects such as RADAR or the European Open Science Cloud (EOSC) show, that no all-encompassing solution has been accepted so far and that there is still a need for further research in this domain.

The PIROL research project will investigate the following research question: How can cross-domain research data publishing processes be facilitated, validated and extended for interdisciplinary reuse? Cross-domain means, that it focuses on interdisciplinary research data management, where data shall be made findable, accessible and reusable over knowledge domain specific boundaries on a meta level, so that researchers and institutions from similar and other disciplines can benefit from it. But it also refers to the nature of domain-specific data silos, and the desire to publish data in an institutional- and platform-independent decentralized fashion.

The rest of the paper discusses the scientific approach to enhance digital research data publishing in a hyper-connected society in the following way: Section 2 describes existing problems and outlines related solutions and the current status of these challenges. Section 3 defines in a concise way the main objectives of the intended PIROL research project. Already existing results and preliminary work from our early stage PhD experience is presented in section 4 before the overall conceptual approach and planned research methodology is defined in section 5. The intended contribution is explicitly listed in section 6 and section 7 summarizes the project scope and limitations.

2 Problem Analysis

We focus on the following three late-breaking challenges in the publishing process for digital research data:

Insufficient data annotation Researchers are not necessarily aware of how to annotate their research data correctly for an interdisciplinary reuse.

Centralized data publishing Researchers commonly have to upload their data on an institutional or even commercial centralized platform.

Weak data validation An assessment of published research data and meta data is often done manually, crowd-based or even not at all. (Semi-)automated reusable processes are still missing.

2.1 Insufficient data annotation

Annotating research artifacts is common and especially in the submission and archiving context of scientific publications well-established. Complex classification systems and taxonomies already exist as well as domain-specific controlled vocabularies with different terms and conventions. The increasing role of knowledge graph in the WWW also introduced alternative approaches to classify and link data. The Semantic Web community proposed a Linked Data approach [1] for that. Unique concept identifiers play an important role and are still under scientific investigation in an interdisciplinary context[1][2][3].

However, to find datasets for a certain research problem is still difficult, especially in an interdisciplinary context. Provided information primarily focuses on discovery meta data, thus provenance information, and less on structured data describing the content of the dataset and other aspects such as access and legal information in a homogeneous fashion.

Furthermore, the annotation process itself seems to be tedious for the majority of users, cognitively overloads them with the large amount of information to provide, and results in sparely provided, narrow-minded meta information with ambiguous content in the end. Ontologies such as schema.org/Dataset can act as a blueprint, but it remains hard especially for inexperienced and non-technical users to select relevant properties and manually annotate a dataset based on these vocabularies; especially in the context of Linked Data. Appropriate components, services and tools in the data management life-cycle could reduce this barrier for a researcher, but it appears that no solution was widely accepted so far, or is bound to a particular data management platform. This especially applies to frontend user interfaces for Linked Data input. Research constantly concentrated on ontology design, but then stopped with the manual creation of meta data information in an RDF serialization format or through simple tabular input interfaces. Only sparse research contributions can be found for frontend components for Linked Data [3] or even a human-centered design approach.

[1] `https://www.project-freya.eu/en/about/mission`
[2] `https://project-thor.eu/`
[3] `https://2019.semantics.cc/robust-identifiers-leading-quality-data`

2.2 Centralized data publishing

Platforms to publish research data encompass institutional-specific infrastructure solutions, Open Data Platforms (CKAN, DuraSpace, Dataverse, EPrints, iRODS, Invenio and others), advanced platforms (Dryad, Zenodo, EUDAT, Figshare) as well as commercial providers such as ResearchGate and Mendeley or even general collaboration and data sharing platforms such as Github, Google Drive, Dropbox and further more [10]. But surveys such as the Wiley Open Science Researcher Insights[4] show, that certain researchers hesitate to use centralized platforms and want to retain control on their research data, especially wrt. data privacy concerns and existing legal agreements. This is reasonable especially in the ongoing debate on data access, protection and data ownership. Decentralized approaches could lower this barrier, increase transparency and eliminate data silos and replicas. Solutions such as e.g., DatProject[5] or EUDAT B2/EOSC[6]) already exist but it appears that no solution was widely accepted so far.

A Linked Data platform (LDP) based approach could even more interconnect research data and its meta data from even heterogeneous platforms in a unique fashion. SoLiD server implementations were already suggested and successfully applied in different scenarios [8], but to the best of our knowledge not tested for research data use cases expect with LDP platform implementations such as Fedora. Further open research challenges exist in how to specify cross-platform access rules and how to run queries among multiple data stores by considering privacy and latency aspects, e.g., by relying on Link Traversal Based Query Execution (LTBQE) [11]

2.3 Weak data validation

High-quality input data is one of several basic requirements for successful business operations because it directly affects consecutive process results in established business value chains. This is especially true for the discovery of published research data in a cross-domain context. However, assuring data quality for published research data and meta data in data repositories is not trivial and often involves human interaction, tedious reviews, or is not done at all.

When research data is published in data repositories together with corresponding meta data information, automatic data conformance and data quality checks can be run. In the past, research primarily focused on reasonable standard metrics that could be used and measured as an indicator for data quality. A survey in 2014 identified 18 appropriate quality dimensions with 69 different metrics from 118 related articles [12] for assessing Linked Data quality. Depending on the type of data source, they allow conformance measurements on data instance level, ontological schema level as well as on service level. It is still a challenge to transfer these data quality metrics to automated data quality assessment processes.

[4] `https://www.wiley.com/network/researchers/licensing-and-open-access/open-science-trends-you-need-to-know-about`

[5] `https://datproject.org/`

[6] `https://www.eosc-hub.eu/`

3 Objectives

To contribute to the systematic, efficient and sustainable cross-domain publication process of research data, PIROL is going to accomplish the following objectives.

O.1 Simplify the annotation process for research data, especially for non-technical users in an interdisciplinary context
O.2 Enable decentralized research data publishing activities on independent platforms for research data
O.3 Support automatic assessment operations on provided research information

4 Preliminary Results

In order to achieve the defined objectives in the problem domain of research data management, expertise in information management, semantic technologies and Web Engineering is needed. The first year of this early-stage PhD project focused on the familiarization with the State of the Art in Linked Data technologies. In the following, we list already achieved preliminary result that will already contribute to the PIROL project. A national growth-core project on Linked Enterprise Data Services[7] in Germany funded by the BMBF supported this initial phase. The primary focus was set on data coherence and data quality assessment operations for general-purpose Linked Data sets accessible via existing web services or placed in a data lake.

A first result was a discussion of definitions on data quality and the deduction of a formalized expression and recommendation on how to compute data quality values for data sources based on multiple criteria on instance, schema and service level (FAME.Q [4]). In a second step, a data quality assessment component for Linked Data sources was implemented that is capable of measuring a basic set of 55 recommended intrinsic, representational, contextual and accessibility quality metrics based on open standards such as W3C's RDF, SPARQL and DQV (SemQuire [7].)A third preliminary result was the design of an ontology that is capable of specifying requirements on data quality characteristics in a uniform fashion to fill a gap between discovery and measurement activities. It can be used for data quality assessment purposes to compare multiple eligible data resources on particular metrics and attributes of current interest. This formalization includes desired measurement result boundaries, a customizable calculation as well as comprehensible quality rating scores (DaQAR [5])

Another research activity was already conducted in the context of building knowledge graphs on interdisciplinary scientific publications. It became obvious, that lots of contributions already exist in the ontological description and backend implementation of Linked Data applications, but that user interfaces require a user to deal with URIs directly. We therefore conducted a user study on how this user interface experience can be improved especially for data input operations and suggested a component-based approach on how to hide Linked Data identifiers based on auto-suggestion features (URI-aware UI [6]).

[7] `http://leds-projekt.de`

5 Approach

Project PIROL: *Publishing Interdisciplinary Research Over Linked data* will extend our preliminary results and focuses on a customer journey in the narrowed problem domain of research data management as depicted in fig. 2

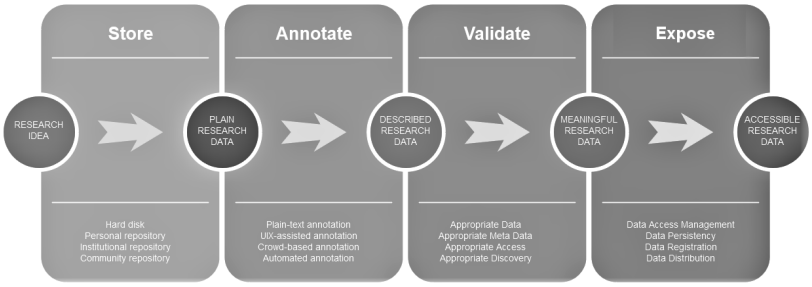

Fig. 2. SAVE Research Data Publishing Journey

We introduce four stages for research data publishing and call them SAVE steps that are aligned to the objectives from section 3. The project will not discuss in detail the Store step with underlying devices and archiving solutions for research data on a physical level. But we will consider persistent versioning aspects and a Linking between derived or related datasets. Future work to enhance these stages is briefly described in the following sections.

5.1 Simplify the annotation process for research data

Research in this section will contribute to the Annotation stage in fig. 2. In a first step, we are going to compare concepts in existing vocabularies on an ontological level with automatic clustering mechanisms. We have the hypothesis that many researchers are not aware of what they should provide as meta information or are not clear which vocabulary to use for annotation purposes. Ontologies such as the schema.org Dataset vocabulary or DataCite are promising approaches, where our approach can contribute recommendations for extension and modification.

In a second step, we investigate appropriate concepts for research data description. Altough Linked Data sources for general-purpose content already exist (examples are DBpedia and Wikidata), this is not necessarily the case for research-specific terms and concepts. Therefore, we will exemplarily introduce an extensible registry for entity URIs to also describe concepts such as a particular measurement method, evaluation method, metric or other characteristic. We will check in a pre-analysis, if existing system platforms such as Wiki-based Linked Data platforms can be used for that purpose.

In a third step, we will conduct research on alternative user input interfaces that assist a user in the process of providing relevant meta data as conceptually shown in fig. 3. This can encompass intelligent data input components, a preprocessing of the provided research data together with an AI-supported classification, as well as context-aware faceted UI approaches.

5.2 Enable decentralized research data publishing activities

Research in this section will contribute to the Exposure stage in fig. 2. In a preparation phase, we will first investigate the current status of existing and established (open) data management platforms, to which extent they support Linked Data references and data privacy aspects. We will then focus on a decentralized LDP approach and apply recent SoLiD considerations on research data management. Following the Linked Data paradigm, a user can place its research data (together with a meta description) on any online platform, where the content is basically accessible and referenceable with a URI.

Therefore, we will develop a concept on how to extend existing platforms with a decentralized SoLiD LDP extension as well as how to run basic queries in this environment. When searching research data with particular characteristics, we are especially interested in how to formulate corresponding queries and how to compare multiple query results. Additionally, approaches on how to actively communicate changes in published research data to research data catalogs in an automated fashion can be discussed.

5.3 Support automatic assessment operations

Research in this section will contribute to the Validation stage in fig. 2. We will put special emphasis on the automated validation operations in the publishing step for research data as we will have a corresponding profile with meta information that can be used for assessment purposes. Post-commit hooks can check, if the published information conforms to predefined requirements.

Furthermore, we want to investigate, which data quality metrics can be applied on research data with Linked Data annotations. Additional data quality metrics for research data may be defined and implemented. This especially applies for metrics that ensure scientific integrity, conformance between the research data itself and the claimed meta information and good scientific practice.

Finally, data quality checks can not only be run in the publication process but for open research data also on data sources in a remote fashion. In combination with the research data discovery and query, we may extend our preliminary work on the definition of data quality assessment requirements (DaQAR).

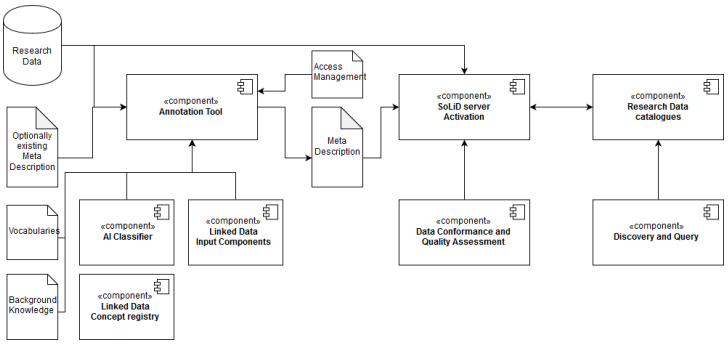

Fig. 3. Basic conceptual architecture

5.4 Research methodology

The PIROL project will focus on a mixed qualitative and quantitative research in information management. It will yield both a concept for systematic and sustainable cross-domain data annotation and publishing activities of research data as well as tested techniques on how to ensure appropriate and gainful cross-domain research data provision from a human factor perspective.

For project preparation purposes, a survey among multiple researchers from different knowledge domains will be conducted to get a basic understanding on the characteristics, content and amount of available research data. New techniques and extensions will then be designed, implemented and tested in an incremental, iterative fashion in small developer teams, that have expertise in Web Engineering, Knowledge Management and Component-oriented Development to develop solutions that are reusable in different contexts. Developed strategies and tools will be evaluated with continuous user questionnaires and lab experiments to which extend they satisfy pre-defined requirements.

6 Contribution

The PIROL research project will contribute a research data publishing process with independent components as depicted in fig. 3 to the problem domain of interdisciplinary and decentralized research data management:

Meta data profile for research data annotation defined by comparing and homogenizing existing vocabularies for interdisciplinary usage

Registry for concept URIs in Linked Research Data established To the best of our knowledge,no similar service exists so far

Research data annotation tool provided To the best of our knowledge, no established platform-independent tool exists so far

Decentralized research data management with SoLiD approved To the best of our knowledge, no Proof-of-concept has been done so far

SoLiD extenders for common data repositories implemented by providing a concept on how to extend traditional data platforms

Query mechanism on SoLiD based research data stores evaluated by applying and extending common query strategies on research data stores

Content of data stores communicated to data catalogs by improving mechanisms to harvest and update (meta)data records with Linked Data

Particular data quality metrics for research data defined by providing algorithms on how to automatically assess relevant aspects

Automated data conformance and quality assessment implemented as data is published without automated check operations so far

Ontology to describe data requirements and characteristics specified by extending existing approaches such as DaQAR and SHACL on how to describe research data characteristics of interest

7 Conclusion

The digital transformation led also in research to new challenges in data management, which includes interdisciplinary annotation and discovery as well as data-privacy issues. The objective of the presented PIROL: *Publishing Interdisciplinary Research On Linked data* is to improve the publishing process of existing (thus persistently stored) referenceable research data for interdisciplinary search and reuse with the means Linked Data. This includes a separate annotation step with a human-centered user input interface concept to create a meta description for concrete research data, the validation of data conformance and data quality aspects between this meta profile and the original research data, and the publishing in an LDP enabled, decentralized infrastructure.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. IJSWIS 5(3), 1–22 (jul 2009)
2. Griffin, P.C., Khadake, J., LeMay, K.S., et al.: Best practice data life cycle approaches for the life sciences. F1000Research 6, 1618 (jun 2018)
3. Khalili, A., Loizou, A., van Harmelen, F.: Adaptive linked data-driven web components: Building flexible and reusable semantic web interfaces. Lecture Notes in Computer Science 9678, 677–692 (2016)
4. Langer, A., Gaedke, M.: Fame.Q - A formal approach to master quality in enterprise linked data. In: Proceedings of the 15th International Conference WWW/Internet 2016. No. October (2016)
5. Langer, A., Gaedke, M.: DaQAR - An ontology for the uniform exchange of comparable LD quality assessment requirements. In: Lecture Notes in Computer Science. vol. 10845 LNCS, pp. 234–242. Springer, Cham (jun 2018)
6. 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)
7. Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: SemQuire - assessing the data quality of linked open data sources based on DQV (2018)
8. Mansour, E., Sambra, A.V., Hawke, S., et al.: A Demonstration of the Solid Platform for Social Web Applications. Proceedings of the 25th International Conference Companion on World Wide Web pp. 223–226 (2016)
9. Sousa, R.B., Cugler, D.C., Malaverri, J.E.G., Medeiros, C.B.: A provenance-based approach to manage long term preservation of scientific data. In: 2014 IEEE 30th ICDE Workshops. pp. 162–133. IEEE (mar 2014)
10. Steinhof, C.: Erfolgskriterien von Forschungsdatenrepositorien und deren Relevanz für verschiedene Stakeholder-Gruppen (2017)
11. Umbrich, J., Hogan, A., Polleres, A., Decker, S.: IOS Press Link Traversal Querying for a Diverse Web of Data. Semantic Web Interoperability, Usability, Applicability 0 (2014)
12. Zaveri, A., Rula, A., Maurino, A., et al.: Quality Assessment for Linked Open Data: A Survey. Semantic Web Journal (by IOS Press) 1, 1–31 (2014)