Connecting the Dots: Transparent FAIRification of Restricted Data

Margherita Martorana[0000−0001−8004−0464]
(Early Stage PhD)

Vrije Universiteit Amsterdam, The Netherlands
m.martorana@vu.nl

Abstract. In the age of information technology and Linked Data, the importance of creating transparent infrastructures for management and exchange of data have emerged to play a focal role for the development of new open research. The FAIR guiding principles have been introduced to advise in the improvement of Findable, Accessible, Interoperable and Reusable technical resources, but the literature is still lacking of practical guidelines for the management and digitization of sensitive and restricted data. The purpose of this research is to introduce new insights aimed at the exploration and discovery of restricted data-sets. The methodology sets out by creating a model for FAIR metadata architecture, which will later be used for automatic ingestion of external restricted data-set files, as well metadata enrichment. Quantitatively and qualitatively evaluations will be performed to assess the workflow, which will later be implemented in the data-set's search engine. Initial results of the metadata ingestion and enrichment are presented as the RML mapper rules and the OPL ontology, respectively.

Keywords: Linked Data, FAIR Principles, RDF Mapping Language, Restricted Data, Metadata Enrichment, Semantic Search, Ontology

1 Introduction

In the era of digitalisation and *Open Science* [15], the management of research and sensitive data is still considered an issue to be solved [7]. Although extensive effort has been made by the scientific community to share and re-use research data, there is still uncertainty on the mechanisms underlying the management of sensitive information. Despite the rapid increase of initiatives and regulations (e.g. GDPR) focusing on this issue, data providers are still lacking technical solutions to allow research groups and institutions to share and re-use their large data banks. One of the most common challenges that is faced by researchers, is the absence of digitized versions of the data. The growing body of information collected by government organisations such as CBS (Statistics Netherlands)[1], is an essential resource for scientists, but it is generally difficult not only to access

[1] https://www.cbs.nl/

but also to find in the first place. In order to support organisations in developing solutions to enable data management and exchange, the FAIR Guiding Principles [17] have been introduced to facilitate reusability and findability of data. The aim is to generate data that is FAIR: Findable, Accessible, Interoperable and Reusable. Central to the entire concept, is the need to integrate both human and machine-readable formats of research data, in order to facilitate the retrieval, analysis and discovery of knowledge.

Many archives in the biomedical domain have already made their data open and findable, such as GenBank [3] and menoci [14], just to mention a few. The latter, together with EUDAT [16], are examples of research data infrastructures based on metadata information. Other projects have proposed certification processes [12] and data sanitization techniques [18] as possible solutions to reproducibility and differential privacy challenges, in the context of confidential research data.

Due to the sensitive nature of a large number of available data-sets, a growing number of initiatives are in facts developing research tools based on the top-level information, the metadata. Nevertheless, formal guidelines for the management and digitization of restricted data and metadata creation is still lacking. Moreover, a broader perspective regarding the use of detailed metadata files as a influential factor in differential privacy and privacy budgets supported queries [19] still need to be explored.

1.1 Restricted Data and Metadata

The term “metadata” is generally understood to mean “data about data”, and it usually describes resources’ embedded information such as authors, dates, versions and technical details [11], but it is regularly seen that important pieces of the puzzle are still missing. In the context of sensitive data, the metadata must be complying with data protection rules, and often important non-confidential information are lacking. For example, the metadata of a certain data-set can report the type of license the researcher has to agree upon for using the data once access is granted. Nevertheless, such information usually reports only the name of the license, and not what the actual agreement defines. For instance, certain license agreements require the researcher to submit their work to the data owner before publishing, and others do not allow for the data to be shown during public speeches or presentations. At this point it is also important to specify the definition of restricted data in the context of this research: by “restricted” we refer to data that is legally bound either by confidentiality or license agreements.

It is clear that such information are truly valuable resources to allow researchers for a more targeted data-set search and exploration, and there is an evident need for such knowledge to be available to the end user.

2 State of the Art

A more detailed account of relevant researches are described in the next section. Firstly, we will discuss the transformation from raw data into Linked Data with the RDF Mapping Language (RML). Afterwards, we will present the Open Digital Rights Language (ODRL) and the CESSDA Metadata Model (CMM), which represent the core frameworks for the novel OPL ontology and the metadata architecture model respectively.

RML. An important step in this research is the creation of Linked Data from unstructured metadata files. For this step, the RDF Mapping Language (RML) [8] has been chosen to map files from data providers, such as DANS, into RDF format. A number of tools based on RML have been developed, and we found that the most user-friendly one is YARRRML [9]. YARRRML is built on top of the RML language, and allows the use of all of its default functions as well as the creation of jar files for custom functions.

ODRL. The Open Digital Rights Language [10] was created with the aim to provide a model and a vocabulary for the expression of statements referring to the usage of services. ODRL has been extensively used thanks to its flexibility and interoperability, and it is now recommended by the W3C as the expression language for describing policies' permissions, prohibitions and constraints. One of the main benefit of using ODRL, is that it allows for the creation of new profiles, grating the community the possibility of expressing additional semantics.

CMM. The Consortium of European Social Science Data Archives (CESSDA) has made extensive efforts in the introduction of various guidelines for data management and maintenance, by highlighting the importance of structured metadata in the social science domain. One of the core activity of the CMO (CESSDA Metadata Office) has been supporting the data-origination process by establishing CMM [5]. The aim of the CMM model is to allow the research community to attain a consistent guide for metadata production, in order to increase consistency and interoperability of this process.

3 Problem Statement and Contributions

The main question of this study is: *How can restricted data be FAIRified?*. This study will focus on social science data, and more specifically on tabular data, but its approach could be extended to any field of research. The assessment of this specific objective consists in the evaluation of the following sub-questions:

RQ1.*How can we represent tabular data that comes with access restrictions in a FAIR manner?* The first step in applying the FAIR principle on restricted research data is the creation of a model or system architecture that is both open and transparent, as well as consistent with the requirements essential to the research community. Due to the constraints imposed by the context of this resource, we will focus on a metadata model that expresses information about the underlying data, therefore avoiding confidential material being exposed.

RQ2.*How can existing scientific data-sets with restricted access be FAIRified in a reproducible and transparent fashion?* The second step consists in the practical evaluation of the approach. In order to achieve this, we aim to map and enrich the metadata of available data-sets, from a raw format into their Linked Data counterparts.

RQ3.***How can researchers dealing with access-restricted data be supported by semantic tools to create data-sets that are FAIR from the start?*** We want to support the research community and data providers in the generation of metadata for restricted tabular data, by creating a transparent tool based on the model proposed. The tool will support the use of semantic web technologies in to order to facilitate logic-based annotations, as well decrease disambiguation in entity-recognition processes.

RQ4.***How can FAIR metadata for access-restricted data improve the effectiveness and usability of a scientific data-set search engine?*** The end goal of this study is to provide a data-set semantic search portal, where users are able to find resources by querying the metadata database. Therefore, our final question for this research, involves the comparison between the performance of the data-set search portal created and other available search environments. Moreover, we will evaluate whether certain implementations in the semantic search of the portal are more beneficial and powerful than others, in order to make suggestions for future work and progress.

3.1 Contributions

The contributions that this research is set to implement are:

- The formulation of a comprehensive metadata model suitable for restricted scientific tabular data, with the input of field experts in the social science domain.
- A method for generating enriched and FAIR metadata by applying the model to existing restricted data-sets.
- The creation of an automated yet transparent and FAIR tool for the generation of Linked Metadata from the upload of data-sets.
- The contribution to a data-set search portal, by implementing semantic search features to optimise the findability of resources.

4 Research Methodology and Approach

The overall approach of this research is based on the principles of: metadata architecture and enrichment, confidentiality-aware data processing, semantic rights representation and access negotiation. Having defined the research questions, we will now address the methodology in more details:

RQ1. A systematic review will be performed to assess, summarise and determine the relevant literature in regards to the management of confidential data-sets, as well as common practices for the exploration and use of restricted data. The outcome of the review will be translated into the metadata model' requirements, and compared to already available resources such as the CMM model, the RDF Discovery Vocabulary (disco) [6] and data management plans' resources [13].

RQ2. Raw metadata can be obtained by different data providers (e.g. CBS and DANS), and can be transformed into Linked Metadata following the model mentioned, using known technologies such as RML. Nevertheless, during this process original information are kept intact, and no extra knowledge is derived. In order to enrich the metadata we can utilise available vocabularies linkable to the original data, such as the DBpedia Ontology [4] and to the European Language Social Science Thesaurus (ELSST) [2]. Moreover, metadata information about data accessibility (e.g. open source, restricted) and data license (e.g. creative common licenses) can be mapped to other known vocabularies such as EuroVoc Thesaurus[2] and the Creative Commons Rights Expression Language (ccREL) [1]. Furthermore, we have created a new Ontology for Policies and Licenses that can also be utilised to map licensing agreements information found in the metadata. In order to map the mentioned vocabularies to the terms in the metadata we can use RML rules, such as the built-in “DBpedia Spotlight Lookup” function.

RQ3. A novel tool is needed to automatize the generation of metadata, therefore supporting restricted-data owners in the creation and release of their resources into digital libraries and metadata portals. Semantic tools, such as entity-recognition and recommendation services will be put in place to map the data-set terms to known thesaurus, such as ELSST, and users will also have the option to customise the terms by adding or suggesting a more appropriate mapping. We aim to create a secure and transparent environment where data owners are in full control over their data, and where the tool workflow is open and accessible to the users, in order to allow them to check every stage of the transformation and decision-making processes.

RQ4. Firstly, semantic search results among different portals and the proposed data-set search engine will be compared, in order to better understand its performance in finding information. Moreover, we aim to analyse the findability of resources, by using two versions of the novel designed portal: 1) this version will be considered to be the “control” group, and it will consist of the portal with all the search functionalities running, 2) this version, instead, will be considered as the “knock-out” group, where certain search functionalities are disabled. The experiment aims to understand which parameters are more powerful and, therefore, to suggested further areas of study. Finally, we will evaluate whether the FAIR implementation in the metadata model allow users to find restricted data-sets, as well as clearly stating the steps and requirements needed to be taken in order to have access to the data from the data-owner. A possible addition to this methodology, is also to investigate privacy budgets supported queries, by tracking users’ history and therefore calculating the level of security of the portal.

[2] http://publications.europa.eu/resource/dataset/eurovoc

5 Evaluation Plan

Having defined the research questions and the methodology of this research, in the following section we will discuss how to evaluate the results.

RQ1. Following the systematic review, metadata requirements found in the literature will be assessed and implemented in the metadata model, and major importance will be given to literature exploring the Reusability and Findability of restricted data. The focus group will be eventually be questioned regarding the final version of the model.

RQ2. A collection of open data-sets in raw format will be used to manually generate their correspondent metadata files. A case study will be performed to quantitatively evaluate the results by counting the number of terms that have been correctly mapped to the model, as well as counting the triples before and after the metadata enrichment. Once the metadata creation is optimised, restricted data-sets from providers such as CBS will be requested and investigated.

RQ3. The tool will be tested by feeding it a collection of data-sets for the automatic mapping to the metadata model, with the aim to achieve the same performance as in the manual mapping mentioned above. Moreover, a focus group consisting of open- and restricted-data owners, as well as researchers, will be asked to perform the task extensively, and their feedback will be assessed.

RQ4. To evaluate the last research questions, we will organise a comparative user study to examine the features implemented in the last iteration of the proposed portal, compared to other available resources. Moreover, we will check whether all requirements found during the systematic review, and further feedback that may arise during this study have been integrated. Furthermore, we will qualitatively assess the strength of semantic search features, by checking the resources returned by the “control” version and the “knock-out” version. Privacy budget queries could also be evaluated by collecting analytical information about the search history of users, and specify a clear privacy cost limit. Finally, the aim of the portal is to include metadata from restricted data-sets, and make them Findable for the research community.

6 Intermediate Results

To create an initial sample for experimentation, RML was used to create a mapping of XML files from the DANSeasy archive, into RDF Linked Data format. RML default functions were used to enrich the metadata, by mapping XML access terms to the OPL ontology. Therefore, we have initials answers to research questions 1 and 2:

- RQ1: during the RDF transformation with RML, the CMM model has been used as a template. We have seen that this model is flexible and expressive enough for initial experiments, but further discussion with the research group are needed to evaluate potential implementations.
- RQ2: metadata enrichment has been performed by mapping XML terms to OPL, and initial feedback shows that a number of additional information

about license agreements and data usage have been added. This step is important in the FAIRification of the model, as the aim is to provide end users with transparent and effective information about the data. The code for the XML mapping can be found at [3], and for OPL ontology at [4].

7 Conclusion

The present study was designed to underline a model for *FAIRification of restricted research data*. Although this research has only started at the beginning of 2021, early findings have significant implications in the first and second research questions, regarding the syntactic and semantic features of the model. In fact, we have shown current metadata models (e.g. CMM) and resources (e.g. ODRL) can be expanded to optimise findability of data. Moreover, early implementations of the OPL ontology have shown how license agreements' details can express important features necessary to the research community. Furthermore, we have acknowledged the usefulness of available tools such as RML during the transformation from raw to structured metadata, and we have also shown how such tools are flexible enough to allow the creation of transparent custom functions to better extract and map information.

Considerable work is needed to gain access to restricted data-sets from providers such as CBS, and a great focus is indeed necessary to guarantee the confidentiality and effectiveness of the model. Notwithstanding these limitation, the aim of this study is to strengthens the importance of metadata and application of the FAIR principles as solutions to restricted data-sets search, and although it focuses on the field of social science, the findings may well have significant implications in other scientific communities.

Acknowledgements. This project is funded by the Netherlands Organisation of Scientific Research (NWO), ODISSEI Roadmap project: 184.035.014. Furthermore, I would like to thank Jacco van Ossenbruggen, Tobias Kuhn and Ronald Siebes for the help and support needed for the success of this research and its future endeavors.

References

1. Hal Abelson, Ben Adida, Mike Linksvayer, and Nathan Yergler. 10. cc rel: The creative commons rights expression language. 2008.
2. Lorna Balkan, Taina Jääskeläinen, Christina Frentzou, and Chryssa Kappi. European language social science thesaurus (elsst): issues in de-signing a multilingual tool for social science researchers. *CHAT 2011: Creation, Harmonization and Application of Terminology Resources*, page 11, 2011.
3. Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.

[3] https://github.com/ritamargherita/opl

[4] https://github.com/ritamargherita/yarrrml-mapper

4. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165, 2009.
5. Kerrin Borschewski, André Förster, Tanja Friedrich, Wolfgang Zenk-Möltgen, Patrícia Miranda, Pedro Moura Ferreira, Jelena Banovic, Aleksandra Bradić-Martinović, Larisa Malic, Henri Ala-Lahti, and et al. Cmm cessda metadata model. Nov 2019.
6. Thomas Bosch, Richard Cyganiak, Arofan Gregory, and Joachim Wackerow. Ddi-rdf discovery vocabulary: a metadata vocabulary for documenting research and survey data. In *LDOW*, 2013.
7. Andrew M Cox, Stephen Pinfield, and Jennifer Smith. Moving a brick building: Uk libraries coping with research data management as a 'wicked'problem. *Journal of Librarianship and Information Science*, 48(1):3–17, 2016.
8. Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. Rdf mapping language (rml). *Specification proposal draft*, 2014.
9. Pieter Heyvaert, Ben De Meester, Anastasia Dimou, and Ruben Verborgh. Declarative rules for linked data generation at your fingertips! In *European Semantic Web Conference*, pages 213–217. Springer, 2018.
10. Renato Ianella. Open digital rights language (odrl). *Open Content Licensing: Cultivating the Creative Commons*, 2007.
11. Elizabeth W King. The ethics of mining for metadata outside of formal discovery. *Penn St. L. Rev.*, 113:801, 2008.
12. Christophe Pérignon, Kamel Gadouche, Christophe Hurlin, Roxane Silberman, and Eric Debonnel. Certify reproducibility with confidential data. *Science*, 365(6449):127–128, 2019.
13. Plato L Smith, Crystal Felima, Fletcher Durant, David Van Kleeck, Hélène Huet, and Laurie N Taylor. Building socio-technical systems to support data management and digital scholarship in the social sciences. In *Anthropological Data in the Digital Age*, pages 31–57. Springer, 2020.
14. M. Suhr, C. Lehmann, C. R. Bauer, T. Bender, C. Knopp, L. Freckmann, B. Öst Hansen, C. Henke, G. Aschenbrandt, L. K. Kühlborn, and et al. Menoci: lightweight extensible web portal enhancing data management for biomedical research projects. *BMC Bioinformatics*, 21(1), Dec 2020.
15. Rubén Vicente-Sáez and Clara Martínez-Fuentes. Open science now: A systematic literature review for an integrated definition. *Journal of business research*, 88:428–436, 2018.
16. Heinrich Widmann and Hannes Thiemann. Eudat b2find: a cross-discipline metadata service and discovery portal. In *EGU General Assembly Conference Abstracts*, pages EPSC2016–8562, 2016.
17. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.
18. Jimmy Ming-Tai Wu, Gautam Srivastava, Alireza Jolfaei, Philippe Fournier-Viger, and Jerry Chun-Wei Lin. Hiding sensitive information in ehealth datasets. *Future Generation Computer Systems*, 117:169–180, 2021.
19. Yang Zhao, Jun Zhao, Jiawen Kang, Zehang Zhang, Dusit Niyato, Shuyu Shi, and Kwok-Yan Lam. A blockchain-based approach for saving and tracking differential-privacy cost. *IEEE Internet of Things Journal*, 8(11):8865–8882, 2021.