Towards a FAIR Sharing of Scientific Experiments: Improving Discoverability and Reusability of Dielectric Measurements of Biological Tissues

Md. Rezaul Karim[1,2], Matthias Heinrichs[2], Lars Christoph Gleim[2], Michael Cochez[1,2,4], Emily Porter[3], Alessandra La Gioia[3], Saqib Salahuddin[3], Martin O'Halloran[3], Stefan Decker[1,2], and Oya Beyan[1,2]

[1] Fraunhofer FIT, Sankt Augustin, Germany
[2] Informatik 5, RWTH Aachen University, Aachen, Germany
[3] Translational Medical Device Laboratory, Lambe Institute of Translational Research, National University of Ireland, Galway, Ireland
[4] Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla, Finland

Abstract. Experiments on the dielectric properties of biological tissues generate data that characterizes the interaction of human tissues with electromagnetic fields. This data is vital for designing electromagnetic-based therapeutic and diagnostic technologies, and for assessing the safety of wireless devices. Despite the importance of the data, poor reporting and lack of metadata impede its reuse and forgo interoperability. Recently, the minimum information model for reporting Dielectric Measurements of Biological Tissues (MINDER) has been developed as a common framework. In this work, we have developed a metadata model and implemented a data sharing framework to improve findability and reproducibility of experimental data inspired by FAIR principles. We define a process for sharing the reported data and present tools to support rich metadata generation based on existing community standards. The developed system is evaluated against competency questions collected from data consumers, and thereby proven to help to interpret and compare data across studies.

Keywords: Scientific Data, Dielectric Measurements, Metadata Management, Semantic Web, FAIR Data Principles.

1 Introduction

Data sharing is the release of research data for use by others [2]. Over the last three decades, there have been many discussions about sharing primary research data. Early studies emphasized the role of sharing scientific data in the practice of open scientific query for verification and refinement of original resources [4]. Within the context of data intensive science, data sharing has become a main vehicle contributing to scientific progress by enabling interdisciplinary interpretation of data, optimizing the use of resources and retaining data integrity for

long term preservation [14]. On the other hand, data driven innovation also requires low barriers to access, interpret and use, rich and widely available data. One of the data intensive innovation areas in health care is the development of medical devices, translating novel research findings to patient care. The design, development, and clinical evaluation of innovative medical devices for diagnostic and therapeutic applications is heavily dependent on findability and reusability of accurate research data.

One of the fastest growing areas for medical device development in Europe is electromagnetic (EM) imaging and therapeutics. Within the context of an aging population and exponential growth in healthcare costs, EM-based techniques provide a very attractive solution for new therapeutics and diagnostic technologies, since they are low cost, non-ionising and largely non-invasive. Many European companies have attempted to commercialize their technology. However, over 75% of medical device companies go out of business within the first five years [5]. Before a new medical device company is formed or a new product line is considered, several factors should be carefully analyzed, such as the clinical need, market size, the regulatory pathway, and, importantly, the technical risk. While many of these factors can be easily quantified, the technical risk often remains the most elusive. Preliminary clinical data or accurate experimental data are often required to de-risk the technical challenge. However, gaps or uncertainty in experimental datasets can mean that the technical risk cannot be estimated sufficiently, and the proposed medical device is ultimately abandoned.

In this work, we aim to bridge the experimental data gap in device development by proposing an approach and implementing tools to improve findability and reusability of dielectric measurement experiments. Achieving well maintained, interoperable, and machine actionable data and metadata are the main building blocks towards this goal. This can be tackled with Semantic Web technologies. The Life Sciences domain, as an early adopter of the Linked Data approach, provides many examples for integrating data from multiple sources and making them queryable on the web through the SPARQL query language [1, 7].

Recently, the FAIR guiding principles have been proposed for scientific data management and stewardship, and have had an impact on the scientific community at large. These principles, rather than prescribing a set of standards, describe the qualities or behaviours required of data sources to achieve their optimal discovery and reuse [10] [18]. The acronym FAIR stands for:

Findable: Enhancing the findability of a given dataset using persistent identifiers while maintaining additional metadata.

Accessible: By using a standardized communication protocol, data, as well as metadata, should be accessible, even if the actual data no longer exists.

Interoperable: Applying controlled vocabularies and qualified references for metadata to be used for knowledge representation.

Reusable: By having a clear and accessible data license and using a well-defined set of accurate vocabulary, the data becomes easily re-usable.

Existing standards can be applied to fulfill the requirements of the FAIR guidelines in varying degrees. However, the need for developing further standards is

apparent. Our specific goal is to adapt some of these FAIR principles using Semantic Web standards to improve discoverability and reusability of experimental dielectric data sets to support EM device development.

We follow an incremental approach to achieve optimal FAIRness of the data sets. This paper reports our first iteration by application of a set of Semantic technologies, and achieves some degree of FAIRness. The current implementation is neither complete in terms of coverage of all FAIR principles nor does it yet demonstrate the full benefits of existing standards. However, this work provides a starting point and a good example of practical applications of Semantic Web technologies for FAIR data sharing, as well as providing a view on future directions.

The rest of the paper is structured as follows: Section 2 gives an overview of related works. In section 3, we discuss the modeling of Dielectric Measurements of Biological Tissues (DMBT) experimental data. Section 4 discusses our proposed pipeline for sharing DMBT data and metadata, domain specific vocabulary development, and our RDFization technique. In section 5, we describe access mechanisms to data and metadata. Finally, we provide a future outlook and conclude the paper.

2 Related Work

Although the knowledge discovery approach was invented more than 400 years ago, the dissemination of knowledge is still mostly done in the same way as it was when invented in the 16th century [3,17]. As published articles are isolated from each other, it is up to the reader to link their information together. This makes information retrieval more difficult than it could be and it is hardly automated.

One of the first attempts to tackle the problem was the *5 star OPEN DATA* by Tim Berners Lee[5]. This approach provides five guidelines for data namely: the data must be available on the web under an open license, must be structured in a well-defined format, accessible in a non-proprietary format like CSV, must use URIs for denotation, and must be linked to other data to add contextual information. Although, some organizations, for example, Thomson-Reuters [13] or Springer-Nature[6] are using Semantic Web approaches to construct knowledge graphs and automating their knowledge retrieval processes, in many other cases still, the data lacks findability, making it hard to access and reuse. Therefore, it is hard to link different datasets, resulting in a lower interoperability. As a complemantary approach the FAIR Data Principles were proposed by the FORCE11 group[7] in 2015 [18] to mitigate these issues as well as to streamline the workflow of scientific data.

[5] http://5stardata.info/en/
[6] https://www.springernature.com/cn/researchers/scigraph
[7] https://www.force11.org/group/fairgroup/fairprinciples

3 Modeling Dielectric Measurements Experimental Data

The dielectric properties, namely, the relative permittivity (ε_r) and conductivity (σ), of biological tissues quantify the interaction of electromagnetic fields with the human body. Together, these properties characterize how EM waves are reflected at, absorbed by, and transmitted through the body. Knowledge of the dielectric properties of various tissues is vital to the field of dosimetry (safety studies, such as for wireless communication devices), and for the implementation of EM-based medical technologies, such as microwave ablation and imaging.

Dielectric measurements are typically performed using an open-ended coaxial probe connected to one port of a vector network analyzer (VNA) through a specialized cable [9]. First, the dielectric probe is calibrated through measurements on materials of known dielectric properties. This enables compensation for systematic measurement errors. Then, the calibration is validated by measuring on yet another known dielectric material, and calculating the accuracy of the measurement. Finally, dielectric measurements of biological tissues are performed by bringing the probe into direct contact with a tissue sample and a dielectric measurement is recorded. After that, the acquired dielectric data may be associated with the material composition within the probe sensing volume. Numerical models may also be fitted to the dielectric data, in order to present the results in closed form.

Although the process of conducting a dielectric measurement on a tissue sample appears rather straightforward, there are a multitude of confounders that can impact the measured data. These confounders are a likely source of inconsistencies in reported data. Both equipment-based measurement confounders and clinical confounders affect the accuracy of dielectric data. Uncertainties in the dielectric data caused by these measurement confounders have been thoroughly investigated over the years and can now be reduced or eliminated by following good measurement practice. However, clinical confounders have been relatively little investigated to date and may introduce a significant level of additional uncertainty into the dielectric data.

The reusability of dielectric measurement data and reproducibility of experiments can be improved by capturing metadata that describes the confounders and by making this metadata a part of the data sharing practice. Moreover, having confounders' metadata together with the metadata about the study itself can help data consumers to define their data requirements and will improve the discoverability of datasets.

Recently as part of our earlier work [12], a set of reporting standards, namely the Minimum Information Model for Dielectric Measurements of Biological Tissues (MINDER) has been proposed [8]. The developed model follows the Investigation-Study-Assay (ISA) framework and defines rich domain metadata to describe the aforementioned clinical confounders such as the tissue source, physiological parameters, in-vivo versus ex-vivo measurement, time, temperature, sample dehydration, as well as dielectric data reporting related confounders such as

[8] https://www.bio-minder.com/

model type selection, number of poles, and fitting algorithms. The developed reporting model is also compliant with the MIRIAM guidelines [8]

In this work, we report on the implementation of a framework to semantically express the metadata and data reported via the MINDER reporting schema and templates. We also developed a platform enabling the discovery of and access to data and metadata by both individuals and machines. To demonstrate the added value of our proposed solution, we collected a set of competency question that are commonly asked by data consumers, as follows:(i) is there any data for pancreas tissue? (ii) is there any data for porcine bladder tissue? (iii) are there kidney measurements available at 24.3 °C? (iv) are there any measurements available on biological tissues at 18 GHz? (v) is there any liver data taken between 20 °C and 25 °C over the frequency range of 1–2 GHz? (vi) is there any tissue-mimicking phantom data? We present that our realized approach enables us to find answers to these questions.

4 Improving the Reusability of Experimental Data

Currently there is no standard way to find and access data on dielectric measurements of biological tissues (DMBT) generated in labs. In most cases, the metadata is only partially recorded, if at all, in lab books and the data is stored separately on hard disks. Some metadata is reported unsystematically in publications, without any controlled vocabulary, resulting in the lack of both human and machine discoverability.

4.1 Pipeline for sharing DMBT data and metadata

In this work, we have created a pipeline to transform the semi-structured, non-standardized DMBT data and metadata resulting from experiments in individual labs to a machine discoverable triple store, and developed a data portal to fulfill data access requirements of end users. Our semantic pipeline implementation follows a subset of the FAIR data principles, namely:

- To make the scientific data *findable*, we assign unique and persistent identifiers to metadata and data, and uploaded it in a publicly accessible repository. We assigned persistent identifiers for each data object. For the identifiers we used dereferenceable URLs through the MIRIAM registry.
- To make our data *accessible*, at first we reviewed the data and metadata and decided on required access mechanisms. Metadata is served via machine interoperable, well defined SPARQL protocol, whereas data can be accessed and downloadable as CSV files with predefined headers. We did not implement any access control mechanism, since all data currently residing on the platform is freely available.
- To make the scientific data *interoperable*, we applied Semantic Web technologies. The metadata is transformed into the Resource Description Framework (RDF) format utilizing shared, domain specific vocabularies and ontologies.

– To make the scientific data *reusable*, we provided rich and well defined metadata. Moreover, reusing existing vocabularies and ontologies makes the shared data linkable to other data sources.

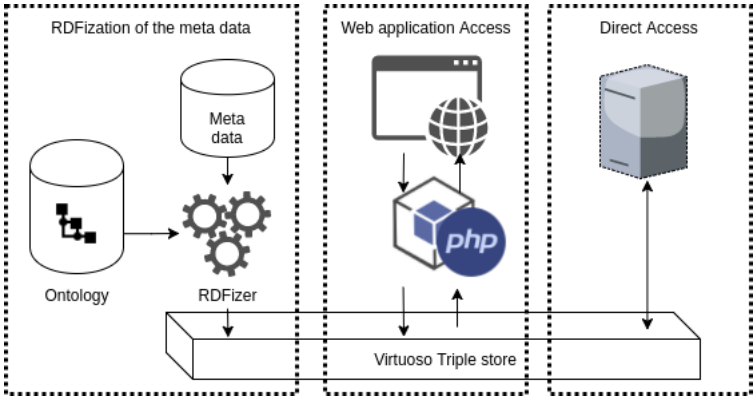

Fig. 1. Proposed architecture used for making the metadata findable and accessible

The basic workflow that results from this work is illustrated in fig. 1. At first, the metadata files are parsed and transformed into Resource Description Framework (RDF) according to the vocabulary. This process is further explained in section 4.3. The resulting triples are then transferred into a Virtuoso triple store[9] that is accessible from the web. The web-application then enables the user to browse this data in various ways, for example, through guided drilling-down into the experiments or the computation of statistics. Further, the user can query the data either directly or using pre-specified query templates. This web application is served using an nginx web server and itself programmed in PHP. Also easy programmable direct access to the data is provided to enable machine access.

4.2 Domain specific vocabulary development

Vocabularies define the terms (concepts and relationships) used to describe and represent a domain. In this work we developed a vocabulary with terms used to define rich experiment metadata. The data model is based on the MINDER minimum information model, which followed the ISA framework classification. In order to optimize interoperability, we reuse a variety of existing terms from established ontologies and vocabularies. Suitable candidates were discovered using the ontology browse and search tools Ontobee [11,20], Linked Open Vocabularies (LOV) [15] and the BioPortal Ontology Recommender [16].

[9] https://virtuoso.openlinksw.com/rdf/

We only reuse terms fully reflecting the semantic meaning of the term in the context of our application and chose definitions from more commonly used ontologies when several suitable candidates were available to simplify integration with existing datasets. This content reuse enables us to develop a consistent representation of this domain, reusing content, deploy existing models and align them to other related datasets. Moreover, this helps us to increase the interoperability between other ontology-based applications.

4.3 RDFization of the experiment metadata

RDF is a W3C standard for the description and modeling of data in a structured way. This standard also provides an abstract and conceptual framework for defining and using metadata and metadata vocabularies by applying statements that consist of subjects, predicates and objects to an ontology. Consequently, the data model becomes easily searchable, findable, and accessible.

The originally generated metadata for our DMBT experiments was stored in Excel format. To transform this data into RDF, we have developed a Java application that parses given metadata files into a data structure, which is then converted into RDF, adhering strictly to the developed vocabulary. The developed *RDFization tool* makes use of the Apache Jena framework. Metadata is stored in an online triple store hosted on Amazon web services. As described above, this endpoint can be accessed directly or through our web application[10]. This web application is designed for non-tech-savvy users to facilate uplaod of new data, however it can receive as input only a specific template.

4.4 Using persistent data and metadata identifiers

To allow reliable referencing of the entities and to enhance the interoperability between different controlled vocabularies and databases, we utilize persistent identifiers (PIDs). PIDs enable the combination of data concerning the same entity from multiple data sources via identifier matching. Thus they enhance the interoperability so that more than one entity can be combined across individual documents and data sources.

The `identifier.org` system directly employs dereferenceable URLs through the MIRIAM registry[11] which provides persistent identification for life science data [6]. Using the MIRIAM Registry service allows for data to be referenced in both a location-independent and a resource-dependent manner, which aids direct resolution of the identifiers via the HTTP protocol [19]. The service's PIDs ensure global uniqueness, perennity, standard compliance, and resolvability while being free of charge to use. The registry is further queryable and features an automated link monitoring system which checks the registered resources on a daily basis for reliability.

[10] https://datalab.rwth-aachen.de/MINDER/
[11] https://www.ebi.ac.uk/miriam/

Overall this system provides a good basis for the persistent identification of data and metadata in our usage scenario and is thus employed as PID provider for all datasets in the Bio-MINDER tissue database.

5 Access to data and metadata through web services

To demonstrate the effectiveness of our proposed approach, we evaluated our approach through our web application. The Semantic Web technologies used in our web application provide the encoding of the entities by assigning resolvable identifiers. The used vocabulary also defines the purpose or interpretation of terms. Additionally, a high-level description of the ontology is available from the web interface so that user can reuse it. These functionalities help make the data accessible and enhances the re-usability.

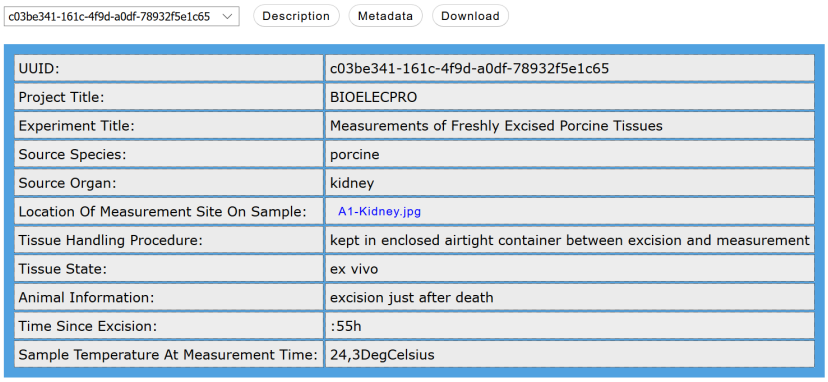

Fig. 2. Metadata for Measurements of Freshly Excised Porcine Tissues

We provided access to data for both human and machine consumption through SPARQL query interface and web application respectively. Also, as part of the web application, there is a tab in which all of the above-mentioned competence questions are answered based on SPARQL queries. These can be executed and the result browsed. Further, when a user selects a specific investigation ID, both the metadata and the description can be accessed. An example is shown in fig. 2. The concrete experimental data can then be downloaded when the user agrees to the specified terms and conditions.

6 Conclusion and Outlook

In this paper, we demonstrated a use case of Semantic Web and a subset of the FAIR principles to improve the repeatability of dielectric measurements of

biological tissue and the reusability of the produced data. We showed how to adopt the MINDER specification and developed a domain specific, controlled vocabulary which makes the data more findable and reusable, setting first steps towards the FAIR principles. We utilized persistent identifiers (PIDs) for both the data and metadata from experiments on the dielectric measurements of biological tissue. This allows the referencing of data in both a location-independent and resource-dependent manner. The provision of resolvable identifiers (URLs) fits well with the Semantic Web vision, and the Linked Data initiative. Moreover, we have reused existing ontologies for terms from our controlled vocabulary where appropriate. Then, we made the overall system available through a web interface which also specifically answers the competence questions which were gathered from domain experts.

In this first iteration, our approach still has several limitations. First, we currently have only met some aspects of the FAIR principles. In future work, we could look at whether it is reasonable and feasible to also address other aspects. For example, we did not deal with licensing properly. We have included a simple licence for the provided files, but these are not in a form which is machine interpretable. One reason we have refrained from defining this is that there is currently no consensus on what a good way would be. This issue is, for example also discussed in working groups on the DCAT standard[12], and several application profiles have chosen different ways to address this. Hence, it would be better to wait until a consensus is reached there before reinventing the wheel ourselves. We have currently also not specified any data access restrictions. For the datasets currently provided, there is no such need, but for data commercially provided, this has to be amended.

Currently, we also chose to only reuse existing vocabulary terms in case there was an exact match. We could consider also adding redundant vocabulary terms, which are broader as the ones we applied to increase potential reusability. A related issue is that we did not make use of, for example, the DCAT vocabulary for specifying our data catalog. The reason is that we were mainly focused on the domain vocabulary and not on the higher level from the start. This can be amended in future work.

This work already showed useful progress. In later work, we would like to further improve this approach for not only biological experiments but also other domain like agriculture, finance, marketing, etc.

References

1. Beyan, O.D., et al.: Querying phenotype-genotype associations across multiple knowledge bases using semantic web technologies. In: Bioinformatics and Bioengineering (BIBE), IEEE 13th International Conference on. pp. 1–5. IEEE (2013)
2. Borgman, C.L.: The conundrum of sharing research data. Journal of the Association for Information Science and Technology 63(6), 1059–1078 (2012)

[12] https://www.w3.org/TR/vocab-dcat/

3. Decker, S.: Rethinking access to scientific knowledge: Knowledge graphs. LinkedIn Pulse (2017), https://www.linkedin.com/pulse/rethinking-scientific-knowledge-graphs-stefan-decker/
4. Fienberg, S.E., Martin, M.E., Straf, M.L.: Sharing research data. National Academy Press (1985)
5. Gage, D.: The Venture Capital Secret: 3 Out of 4 Start-Ups Fail the wall street journal (2012), http://www.wsj.com/articles/SB10000872396390443720204578004980476429190
6. Juty, N., Le Novère, N., Laibe, C.: Identifiers. org and MIRIAM registry: community resources to provide persistent identification. Nucleic acids research 40(D1), D580–D586 (2011)
7. Kazemzadeh, L., Kamdar, M.R., et al.: LinkedPPI: enabling intuitive, integrative protein-protein interaction discovery. In: Proceedings of the 4th International Conference on Linked Science-Volume 1282. pp. 48–59. CEUR-WS. org (2014)
8. Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., et al.: Minimum information requested in the annotation of biochemical models (MIRIAM). Nature biotechnology 23(12), 1509 (2005)
9. Meaney, P.M., Gregory, A.P., Seppälä, J., Lahtinen, T.: Open-ended coaxial dielectric probe effective penetration depth determination. IEEE transactions on microwave theory and techniques 64(3), 915–923 (2016)
10. Mons, B., Neylon, C., Velterop, J., Dumontier, M., et al.: Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the european open science cloud. Information Services & Use (Preprint), 1–8 (2017)
11. Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., et al.: Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic acids research 45(D1), D347–D352 (2016)
12. Porter, E., La Gioia, A., Salahuddin, S., Decker, S., Shahzad, A., et al.: Minimum information for dielectric measurements of biological tissues (MINDER): A framework for repeatable and reusable data. International Journal of RF and Microwave Computer-Aided Engineering pp. e21201–n/a, e21201
13. Song, D., Schilder, F., Hertz, S., Saltini, G., et al.: Building and querying an enterprise knowledge graph. IEEE Transactions on Services Computing (2017)
14. Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M.: Data sharing by scientists: practices and perceptions. PloS one 6(6), e21101 (2011)
15. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semantic Web 8(3), 437–452 (2017)
16. Whetzel, P.L., Noy, N.F., et al.: BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research 39(suppl_2), W541–W545 (2011)
17. Whewell, W.: History of inductive sciences (1858)
18. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data 3, 160018 (2016)
19. Wimalaratne, S.M., Bolleman, J., Juty, N., Katayama, T., Dumontier, M., Redaschi, N., Le Novère, N., othersection 4.2?: SPARQL-enabled identifier conversion with identifiers.org. Bioinformatics 31(11), 1875–1877 (2015)
20. Xiang, Z., Mungall, C., Ruttenberg, A., He, Y.: Ontobee: A linked data server and browser for ontology terms. In: ICBO (2011)