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

Md. Rezaul Karim\textsuperscript{1,2}, Matthias Heinrichs\textsuperscript{2}, Lars Christoph Gleim\textsuperscript{2}, Michael Cochez\textsuperscript{1,2,4}, Emily Porter\textsuperscript{3}, Alessandra La Gioia\textsuperscript{3}, Sahib Salahuddin\textsuperscript{3}, Martin O’Halloran\textsuperscript{3}, Stefan Decker\textsuperscript{1,2}, and Oya Beyan\textsuperscript{1,2}

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

Abstract. The dielectric properties of biological tissue experiments generate data characterizing the interaction of human tissues with electric fields. This data is vital for designing electromagnetic-based therapeutic and diagnostic techniques, and for assessing the safety of wireless devices. Despite the importance of the data, poor reporting and lack of metadata impede its reuse 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 implement 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 that way 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 \cite{2}. During the last three decades, there were 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 \cite{3}. 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 \cite{11}. 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 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 attempt to commercialize their technology. However, over 75\% of medical device companies go out of business within the first five years \cite{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 \cite{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 \cite{9}. 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 data set 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 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 demonstrates the full benefits of existing standards. However, it 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 mechanism 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 has been invented in the 16th century [3,14]. 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.

Nevertheless, many organizations, for example, Thomson-Reuters\footnote{https://www.thomsonreuters.com/en.html} or Springer-Nature\footnote{http://www.springernature.com/de/} are using Semantic Web approaches to construct knowledge graphs and automating their knowledge retrieval processes. As a lot of research data must be managed, approaches exist attempting to solve this problem. One of them is the 5 star OPEN DATA proposal by Tim Berners Lee\footnote{http://5stardata.info/en/}. As the name suggests, the proposal puts forward five guidelines for data:

1. The data must be available on the web under an open license
2. The data must be structured in a well-defined format
3. The data must be accessible in a non-proprietary format like CSV
4. The data must use URIs for denotation
5. The data must be linked to other data to add contextual information.

Reflecting the current state-of-the-art, this approach shows a modern understanding of the Semantic Web vision. Generally, if one applies these five principles to a given set of data, the data will be well managed. At first glance, this may sound sufficient, but if one takes a closer look at the limitations of approaches like the given one become obvious.

Due to the approach only focusing on data, whereas ignoring existing metadata, only the information itself is well-formatted and readable. 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. To mitigate these issues as well as to streamline the workflow of scientific data, the FAIR Data Principles were proposed by the FORCE11 group\footnote{https://www.force11.org/group/fairgroup/fairprinciples} in 2015\footnote{https://www.force11.org/group/fairgroup/fairprinciples}. The acronym FAIR stands for:

5 https://www.thomsonreuters.com/en.html
6 http://www.springernature.com/de/
7 http://5stardata.info/en/
8 https://www.force11.org/group/fairgroup/fairprinciples
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.

3 Modeling Dielectric Measurements Experimental Data

The dielectric properties, namely, the relative permittivity ($\varepsilon_r$) and conductivity ($\sigma$), 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. These experiments simulate in vivo scenarios where the system is perturbed by bringing the probe in contact with the patient’s body. The system perturbation is introduced by moving and changing the cable position. Then, an ECal module is used to automatically refresh the system calibration after the occurrence of a perturbation that introduces noise into the dielectric data. After that, the acquired dielectric data is associated with the material composition within the probe sensing volume. The experimental outcome is validated and integrated with numerical results obtained by probe-material simulations. The integrated data provides information on how to interpret the dielectric properties of heterogeneous tissues, thus allowing refinement of the measurement protocol of tissue dielectric properties.

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 the main 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 and reproducibility of experiments can be improved by capturing metadata that describes 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, a set of reporting standards, namely the Minimum Information Model for Dielectric Measurements of Biological Tissues (MINDER) has been proposed by the Bio-MINDER team\textsuperscript{10}. 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 and ex-vivo measurement, time, temperature, sample dehydration, as well as dielectric data reporting related confounders such as model type selection, number of poles, and fitting algorithms. The developed reporting model is also compliant with the MIRIAM guidelines \textsuperscript{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 questions that are commonly asked by data consumers, as follows:

1. Is there any data for pancreas tissue?
2. Is there any data for porcine bladder tissue?
3. Are there kidney measurements available at 24.3°C?
4. Are there any measurements available on biological tissues at 18 GHz?
5. Is there any liver data taken between 20°C and 25°C over the frequency range of 1–2 GHz?
6. Is there any tissue-mimicking phantom data?

We will show 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 dielectric measurements of biological tissues (DMBT) data 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

\textsuperscript{10} https://www.bio-minder.com/
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 assign the 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.

![Diagram](image)

**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 ontology. This process is further explained in section 4.3. The resulting triples are then transferred into a Virtuoso triple store[^11] 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 [10,17], Linked Open Vocabularies (LOV) [12] and the BioPortal Ontology Recommender [13] (see also appendix A).

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 increasing the interoperability between other ontology-based applications. A high level overview of the developed ontology can be found in figure 2. The ontology is conceptually split into three abstraction levels (investigation, experiment and assay) each of which is in itself a meaningful ontology.

![Fig. 2. High-level view of the dielectric measurements ontology](image-url)
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 ontology. The so-called RDFization tool makes use of the Apache Jena framework\(\text{12}\).

Now, to make the metadata accessible and retrievable from the web, we stored them in an online triple store\(\text{13}\) hosted on Amazon web services. As described above, this endpoint can be accessed directly or through our web application\(\text{14}\). This web application is created to increase the findability and accessibility of the data and metadata, also for non-tech-savvy users.

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 utilized persistent identifiers (PIDs). To achieve this we rely on the identifier.org\(\text{15}\) system which employs directly dereferenceable URLs through the MIRIAM registry\(\text{16}\) that provides persistent identification for life science data \(\text{[6]}\). Using the MIRIAM Registry service allows us to reference data in both a location-independent and a resource-dependent manner, which aids direct resolution of the identifiers via the HTTP protocol \(\text{[16]}\).

Using PIDs, it is further possible to combine data concerning the same entity from multiple data sources via identifier matching. Using PIDs enhances the interoperability so that more than one entities can still be combined across individual documents and data sources. The PIDs assigned by Identifiers.org 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. Overall this system provides a good basis for the persistent identification of data and metadata in our usage scenario.

\(\text{12}\) https://jena.apache.org/
\(\text{13}\) http://ec2-54-77-195-222.eu-west-1.compute.amazonaws.com:8890/sparql
\(\text{14}\) https://datalab.rwth-aachen.de/MINDER/
\(\text{15}\) http://identifiers.org/
\(\text{16}\) https://www.ebi.ac.uk/miriam/
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. Besides, 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.

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 the above mentioned competence questions are answered based on SPARQL queries. These can be executed and the result browsed. Further, when a user selects specific investigation ID, both the metadata and the description can be accessed. An example is shown in fig. 3. 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 experiments 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 (PIIDs) for both the data and metadata from the dielectric measurements of biological tissue.
experiments. 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 some terms from our controlled vocabulary. Then, we made the overall system available through a web interface which also specifically answers the competence questions which were gathered from the domain experts.

In this first iteration, our approach still has several limitations. First, we currently only reach some aspects of the FAIR principles. In future work, we could look 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\(^{17}\) 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


\(^{17}\) https://www.w3.org/TR/vocab-dcat/
A  Examples of Vocabulary Reuse
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement</td>
<td>Frequency (units)</td>
<td>Hz, kHz, MHz, GHz, THz</td>
</tr>
<tr>
<td>Temperature</td>
<td>Accuracy (units)</td>
<td>C (Celsius), F (Fahrenheit), K (Kelvin)</td>
</tr>
<tr>
<td>Timing</td>
<td>Accuracy (units)</td>
<td>h, min, s, less than (min), less than (h), more than (min), more than (h)</td>
</tr>
<tr>
<td>Measurement</td>
<td>Power (units)</td>
<td>dBm, mW, NA</td>
</tr>
<tr>
<td>Tissue state</td>
<td>Ex-vivo, in-vivo, in-vitro</td>
<td></td>
</tr>
<tr>
<td>Validation material</td>
<td>Butanol, ethanediol, deionised water, methanol, ethanol, 0.1 M NaCl, 0.9% NaCl, other</td>
<td></td>
</tr>
<tr>
<td>Type of sample</td>
<td>Liquid, tissue, other</td>
<td></td>
</tr>
<tr>
<td>Tissue source</td>
<td>Organism</td>
<td></td>
</tr>
<tr>
<td>Tissue source</td>
<td>Type</td>
<td>Lung, liver, muscle, kidney, heart, breast, skin, adrenal,</td>
</tr>
<tr>
<td>Tissue diseased</td>
<td>Diseased, normal, unknown</td>
<td></td>
</tr>
<tr>
<td>Liquid volume</td>
<td>Units</td>
<td>uL (microlitre), mL (millilitre), cL (centilitre), dL (decilitre), L (litre)</td>
</tr>
<tr>
<td>Measurement error</td>
<td>Type</td>
<td>Accuracy, repeatability, TCU (total combined uncertainty), other</td>
</tr>
</tbody>
</table>