Fieldlab Report and Team Structure

Xinyue Xie, Yuan Lin, Wei Chen, Yiqin Lei, Yijie Lu

1. Introduction

In this report, we first give an introduction about our motivation and topic, and then, make a problem description in the second section. We also mentioned the task and the information we need in this section, which might be related to the content about the interview we are going to have.

In the third section, we briefly reviewed some content related to this problem. Some concepts, frameworks, current techniques and some related work done by other researchers are explained in this part.

The fourth section is about our team. We make a brief introduction about ourselves, our background, our competencies, and therefore, our task division of each teammate. Apart from that, this section also contain an import part, our stakeholders and their contact information. They are whom we want to have an interview with, and whom we will ask for some technical support.

The last section is about our future plan. It contains how we will begin this task and, recently, what information we hope to get from the interview. We list some topics there, hoping that this will make our interview progress more smoothly.

2. Problem Statement

The problem here is to develop metadata templates (creation in human and machine-readable format) based on rigorous controlled vocabulary and ontology specification, that is, a set of commonly agreed terms in a domain, for a clinical research on a new vaccine developed in LUMC, and for data from OPD (Outpatient department) and MCC (Mother Childcare) in Zimbabwe.

To develop those meta-data templates, understanding all aspects of the data is necessary. For instance,

- The source of data.

- When and how is the data collected?

- What information the data have?

- Are there any exist templates, and if so, what are the differences between the templates for different regions, and what information or meta-data the templates have?

- Who work with them?

- What medical expertise is there to know?

In addition, in order for the metadata templates to describe the data more comprehensively (making metadata "rich"), knowledge or terminology from medical or other fields is needed, which requires extensive collaboration with experts in related fields. It also help to meet domain-relevant community metadata standards.

Finally, create good meta-data quality that require reporting guidelines to provide a standardized structure for the metadata components, ontologies mentioned above and technology to make it easy to author good metadata in the first place. By doing so to implement FAIR principles.

3. Current Techniques and Related Work

3.1. FAIR

When designing vaccine metadata templates, we need to implement the FAIR principle, which provides guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets.

For the Findable principle, the first and most important point is that we need to assign a unique identifier to each element of data and metadata, and this identifier must be sustainable so that data users can accurately find the data. Secondly, the data needs to be described by rich metadata, so that even if there is no data identifier, the data can be found based on the information provided by the metadata. Metadata includes "intrinsic metadata" used to describe some automatically captured and "contextual metadata" used to describe the surrounding environment of the data. The third point is that the association between metadata files and data should be clarified by the unique identifier of the data mentioned in the metadata. The fourth point is that data and metadata need to be registered or indexed to some resources that can be searched.

For the Accessibility principle, it mainly focuses on allowing data and metadata to be accessed by their unique identifiers, so avoid using some special tools or communication methods. The agreement should be free and open, and can be implemented on a global scale, thereby facilitating data retrieval. But at the same time, verification and authorization

are needed when necessary to ensure that some relatively sensitive private information can also be FAIR. The second point is that the storage cost of metadata is relatively low, so even if the data degrades or disappears over time, the metadata should be accessible and continue to exist.

For Interoperability, first, data and metadata need to use a formal, accessible, shared, and broadly applicable language for knowledge representation. Common controlled vocabulary, ontology, and the previously mentioned unique identifier need to be used, and a good data model is required. A good framework to describe data and metadata. The second point is that the controlled vocabulary used to describe the data set needs to use a globally unique and sustainable identifier, so that anyone who uses the data set can easily find and access these documents. The third point is that data and metadata should be cross-referenced with other data to establish as many meaningful connections as possible, thereby enriching the context of data and metadata.

For the Reuse principle, it is related to the second point in Findability. The data must have a variety of accurate and related attributes. This kind of data with many additional tags will make it easier to reuse the data. Reuseble pays more attention to the user's ability to determine whether data is useful in a specified environment. Therefore, in this principle, it is necessary not only to provide metadata to facilitate the discovery and search of data, but also to provide metadata about the background of data generation, such as experimental protocols and data creation. The manufacturer, brand, and medicine used of the machine or sensor. When providing these, the data publisher should not guess the field used by the data but should provide the information as fair and comprehensive as possible. Data and metadata should clearly describe the usage rights of the data, so that it can be more reusable in law. The data also needs to be related to its detailed source, and it must be traceable. At the same time, the data must comply with regional standards, organize the data in a standardized way, use mature and sustainable file formats, follow common templates, and use metadata with strict vocabulary control.

3.2. Metadata

Metadata is the data that store attribute of a certain data, to explain in one sentence, it is data about data. There are many types of metadata, Descriptive metadata, Structural metadata, administrative metadata, Reference metadata, Statistical metadata and Legal metadata.

For example, Descriptive metadata describe the attribute to identity data, like title, abstract, author and keywords. Take a book for example, the book itself is the data and the information about the book is metadata. When finding a book from the website or in bookshop, the more information you know about the book (metadata), like author name,

publisher, publish date, etc., the easier you find the book.

Therefore, it is important for the metadata to be correct and rich. Knowing only one attribute, like the name of the book, the author, the publisher is not enough for finding the book. Add more information can narrow down the searching target. The next problem is to decide how rich the metadata need to be. The range depend on the purpose of the data. One principle of the FAIR data is to be reusable. This can be challenge for collecting metadata because the metadata may be diverse from the purpose of the project.

3.3. Related Work

In Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. They introduce how to apply fair principle to hospital data. The research use an ontological model to transfer data and metadata into a FAIR based architecture.

This FAIR data is prepared to be reused for federated analysis, linkable to other FAIR data such as Linked Open Data, and reusable to develop software applications on top of them for hypothesis generation and knowledge discovery.

4. Team Structure

Aiming to explore more about the fieldlab, our team consists of 5 students: Xinyue Xie, who was major in *Automation*, Wei chen was major in *Software Engineering*, Yuan Lin and Yijie Lu were major in *Information Management*, Yiqin Lei was major in *Information & Computing Science*. Despite coming from different disciplines, we are currently all in the *Data Science of Computer Science* program.

- Xinyue has considerable proficiency in process design, which helps us obtain guidelines of reporting FAIR data and metadata.

- Wei is fairly proficient in coding and approaching tasks by programs, making things more efficient.

- Besides, Yuan and Yijie can grasp the key information from original data set, contributing to obtain FAIR data and metadata.

- In addition, Yiqin can formulate a model for the original data to judge whether it can be FAIR or metadata, reducing the repeat work.

Regarding the filedlab as a project, we need to communicate with mentors to get suggestions from them. As we separate the whole filedlab into to two parts, one is to form a metadata template about diarrhea vaccine which is developed by LUMC and has been used in Zambia and Burkina Fasso, the another is to form a template for Zimbabwe based on

the data from the combination of OPR (Outpatient Registry) and MCC (Mother & Childcare).

In this way, we have got the stakeholders of the former to interview :

- Prof. Meta Roestenberg,
  ......@...

- Jacqueline,
  ......@...

- Aliya Aktau,
  aleka.aktau@gmail.com

- Mariam Basaiic,
  mariam.basajja@gmail.com

Relatively, we have four mentors in contact for the latter:

- Aliya Aktau,
  aleka.aktau@gmail.com

- Mariam Basaiic,
  mariam.basajja@gmail.com

- Ronald Manhibi,
  ......@...

- Kudakwashe Madzene,
  ......@...

The missing contact information would be filled after our conformation.

They are professional in this field and we will get in touch with them for further work. Aiming to practice our ability of communication and presentation, we decided to be the ones to contact mentors and organize the meeting respectively each week.

Prof. Meta Roestenberg and data scientist Jacqueline are in charge of the first part about the vaccine data in Zambia and Burkina. Actually, the data perhaps is distributed in different formats in Zambia and Burkina, for which we should build a template for it in LUMC. After clarifying the task, we will try to contact with the two experts.

For the part of Zimbabwe, we will try to have meetings with Mariam in LUMC and Aliya via Zoom. Ronald Manhibi and Kudakwashe Madzene are the experts who are in charge of this project. In this part, the data consists of OPR/OPD and MCC in Zimbabwe may meet some questions which we will ask experts in the future interview.

In terms of leadership of our team, we considered that it depends on the small task of the project, in other words, everyone is the leader when the task is in his/her major. And

everyone needs to present clearly about his/her work to each other in order to make a proper skeleton. Yiqin and Yijie would check the logical aspects and data structure. Xinyue and Yuan would evaluate the details of each part if possible.

For little knowledge about the proper things of our fieldlab, the structure will be adapted to future work.

5. Preparation

Further work needs to be done to figure out the problems we obtained. As we are aiming to obtain the metadata template of the new vaccine developed by LUMC, we need to grasp what is the vaccine about, how does it comes and many other question.

In this way, we have been learning for conceptions and the principle of metadata and FAIR data as above. But the more vital thing is to communicate with the mentors, for which we plan to organize a meeting this week.

Biology knowledge is also important for us to understand and evaluate the ontology of FAIR data. To make progress on this aspect that we are lack, a seminar would be held once a week if possible in such case we know what to learn about the vaccine.

If everything goes in a proper way, we will gradually grasp the method to obtain the metadata template of the new vaccine which follows the FAIR principles with the help of mentors.

5.1. Interview

After our first meeting with Aliya, our mentor, the purpose of this Fieldlab has become more clear. We plan to interview the stakeholders from the research institute in Zimbabwe, Zambia, Burkina Faso and LUMC.

For stakeholders from Zambia, Burkina Faso and LUMC, we want to know that,

1. What data the they have, for example, more information of Pre-vaccination and Post-vaccination, as well as the record of infant.

2. What vaccine do they have, the type of vaccine and what disease are the vaccine used for.

3. Is there any exist templates, what is the advantages and disadvantages of it.

4. Who work with them, research institute, other medical institutions, etc.

5. Where is the data collected and stored.

Beside this, we will also make the template for the hospital in Zimbabwe. For this part, since the hospital over there

165 may not have a complete computer system, and there are
166 already some paper records in use in MCD and MCC. So
167 during the interview, we would like to get to know more
168 detailed from the exist data form, for instance, what record
169 OPD and MCC have, and do they have a current metadata
170 template in use.