Integrated Citizen Centered Digital Health and Social Care
A. Värri et al. (Eds.)

doi:10.3233/SHTI200690

Security and Privacy when Applying FAIR Principles to Genomic Information

Jaime DELGADO[1] and Silvia LLORENTE
Information Modeling and Processing (IMP) group – DMAG,
Computer Architecture Dept. (DAC),
Universitat Politècnica de Catalunya (UPC BarcelonaTECH)

Abstract. Making data Findable, Accessible, Interoperable and Reusable (FAIR) is a good approach when data needs to be shared. However, security and privacy are still critical aspects. In the FAIRification process, there is a need both for de-identification of data and for license attribution. The paper analyses some of the issues related to this process when the objective is sharing genomic information. The main results are the identification of the already existing standards that could be used for this purpose and how to combine them. Nevertheless, the area is quickly evolving and more specific standards could be specified.

Keywords. FAIR, FAIRification, de-identification, anonymization, license attribution, privacy, rules, genomics

1. Introduction

The FAIR data principles consist on making data Findable, Accessible, Interoperable and Reusable. They were first formally introduced in [1]. When data (very often "scientific data") is to be made publicly available, even subject to some conditions, a good approach is to achieve these principles. The process by which data is converted or adapted to be FAIR is very often called FAIRification. There are many aspects to be considered when FAIRifying data. This paper focuses in the security and privacy aspects. In addition, we also focus on a specific kind of data: health data, including genomic data.

Part of this work has been done in the context of the FAIR4Health European Project [2], which provides real scenarios where to apply FAIR principles and privacy aspects.

The Methods section analyses the FAIRification process and its impact in security and privacy, while section 3 on Results provides details on the available international standards dealing with the de-identification, anonymization and pseudonymization issues. On the other hand, the Discussion concentrates on the License attribution step and all the related problems that need to be solved. Finally, the Conclusions point to some more ideas on future work.

[1] Corresponding Author, Jaime Delgado, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain; e-mail: jaime.delgado@upc.edu

2. Methods – FAIR concepts

The FAIR principles need to be applied, through a process, to have health information Findable, Accessible, Interoperable and Reusable. This FAIRification process consists of a set of steps that need to be followed to prepare the data.

There are several initiatives for the specification of the FAIR workflow or FAIRification process. Moreover, there are different definitions for those processes, although most of the approaches are very similar.

For example, GO FAIR, an initiative that aims to implement the FAIR data principles, specifies its own FAIRification process [3]. They propose guidelines to help in making the data FAIR.

On the other hand, the FAIR4Health project [2] has developed its own workflow based on the FAIRification process adopted by GO FAIR. The FAIR4Health specification is the starting point for our analysis.

The different steps of the FAIR4Health's FAIRification workflow could be summarized as: 1) Raw data analysis, 2) Data curation & validation, 3) Data de-identification / anonymization, 4) Semantic modeling, 5) Make data linkable, 6) License attribution, 7) Data versioning, 8) (Meta)data aggregation, and 9) Archiving.

A first consideration of these steps from a Security and Privacy (S&P) point of view leads to the issues described in Section 3 on Results. The relevant steps are "Data de-identification / anonymization" (step 3) and "License attribution" (step 6).

If we compare with the GO FAIR initiative, they also define a step on licenses (called "Assign license"), making clear that, "although license information is part of the metadata, they have incorporated the license assignment as a separate step in the FAIRification process to highlight its importance". It is very important to take into account that in many situations having a license is the only way to access the data.

The Research Data Alliance [4] is also very active in the area. Together with FORCE11 [5], they have jointly created the FAIRsharing.org registry of standards and other resources [6]. The registry collects metadata to ensure that the information is FAIR, claiming that one way to achieve accessibility (the "A" from "FAIR") might be "by identifying their level of openness and/or license type".

Finally, in relation to the S&P aspects, GO FAIR refines the 4 principles. For example, with A1.2 (The protocol allows for an authentication and authorization where necessary) and R1.1 ((Meta)data are released with a clear and accessible data usage license). From this, the Research Data Alliance identifies the importance of the evaluation of the fulfillment of these principles, what they call the "FAIR Data Maturity Model". In the S&P identified aspects, it means that data providers should evaluate if the access protocol supports authentication and authorization and if metadata refers to a standard license.

3. Results - Analysis of Security and Privacy aspects

The first results of our work are an analysis of the S&P relevant FAIRification steps previously identified. Specifically, de-identification, pseudonymization, anonymization and license attribution.

3.1. De-identification, anonymization and pseudonymization

Data de-identification/anonymization, step 3 of the FAIRification process, is the first step that explicitly refers to S&P. It recommends applying de-identification, anonymization or both operations to the dataset with the objective of enabling data sharing without compromising data subjects' rights regarding privacy issues.

For de-identification, the simplest approach is to drop data elements from the dataset. However, different understandings of the terminology for these concepts should be taken into account, as those from ISO/IEC 20889:2018 (Privacy enhancing data de-identification terminology and classification of techniques) [7].

In addition, ISO 25237:2017 on Pseudonymization [8] introduces several definitions to understand the relationship between the concepts of "de-identification", "anonymization" and "pseudonymization".

In particular, anonymization is understood as the "process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party". However, there is a very relevant note to this definition clarifying that "the concept is absolute, and in practice, it may be difficult to obtain". Therefore, anonymized data could be still considered as personal data if it is not possible to guarantee the absolute impossibility of re-identifying the data. On the contrary, it would no longer be personal data, so there would be no need to comply with the data protection requirements.

Next, de-identification is defined as a "general term for any process of reducing the association between a set of identifying data and the data subject", and pseudonymization as a "particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms". A trusted third party may be able to obtain the normal personal identifier from the pseudonym.

There is no specific ISO standard on anonymization. However, ISO/IEC 20889:2018 [7], introduced before, focuses on commonly used techniques for de-identification of structured datasets as well as on datasets containing information about data principals.

The use of de-identification techniques is good practice to mitigate re-identification risk, but does not always guarantee the desired result. This de-identification standard [7] "establishes the notion of a formal privacy measurement model as an approach to the application of data de-identification techniques". In any case, the application of these techniques should be considered as a privacy risk in the Privacy Impact Assessment.

3.2. License attribution

License attribution, step number 6, is the second FAIRification step that refers to S&P.

The objective of this step is to make clear the need for a regulatory framework for data owners to provide licensing attributions. The purpose of licenses is to support the proper reusability. Although use of Creative Commons [9] is a possible approach, other licensing options might be considered, since there might be very different needs for research datasets including health or genomic data.

As the FAIR4Health project states, the license attribution for the dataset should be always clearly stated, together with the process by which an external requester could demand the permission for reusing the dataset. It should be also taken into account, as mentioned before, that the absence of an explicit license may prevent others to reuse data, even if the data is intended to be open access.

The previous issues raise the fact that there are additional problems to consider when licenses are in place, as developed in section 4.

4. Discussion

Our discussion focuses on proposing solutions to help implementing the step 6 of the FAIRification process; i.e. "License attribution". The related issues include:

- How to express the licenses.
- How to protect them and guarantee their provenance.
- How to evaluate their authorization.
- How to enforce what they are controlling.

The proposed approach is based on the idea of access authorization using privacy rules, which describe the conditions for accessing the information, including allowed actions, analysis purposes or algorithms. It is also very important to support different levels of granularity in the allowed access to the information.

A second focus on the consideration of these potential problems on license management, is the selection of a specific type of information with high privacy requirements: genomic information. There are different ways and standards to represent this kind of information. For our analysis, we start with MPEG-G [10], an ISO Standard for the representation of genomic information. We do not consider this as a limitation since MPEG-G already integrates different aspects of security and privacy, which could be used for our purposes. If we handle genomic information in different formats, we still would have very similar S&P issues.

Regarding license expression (our first issue) and protection and provenance (the second one), MPEG-G, in its part 3 [10] provides an access control mechanism based on privacy rules, exactly as we are proposing. These rules are expressed in XACML [11], a general purpose language for access control rules definition. It allows a high level of granularity, which is very convenient for our case. The rules (that are in fact metadata) are included in the genomic information structure to be protected, and an authorization mechanism is also defined in the standard, based on the genomic file structure and the hierarchy of elements inside it. Privacy rules are located inside special protection elements associated to different kinds of genomic information (and also metadata) inside the file. MPEG-G defines mechanisms to ensure rules integrity, like digital signatures associated to them. Provenance can be checked from these signatures. Moreover, protection elements may contain encryption parameters for protecting both the genomic file and its metadata, also providing the required protection.

Finally, authorization and enforcement mechanisms are also considered in MPEG-G. [12] graphically explains how MPEG-G authorization works based on the hierarchical file structure, which can represent from several complete genomic studies to the more basic data units. Enforcement is guaranteed by the information described in the rule. Only the actions defined inside the rule over the corresponding data will be allowed by the authorization process.

To sum up, MPEG-G is a suitable example of how license related issues can be solved when trying to apply FAIR principles to genomic information.

5. Conclusions

This paper has presented the issues to consider when providing security and privacy in the process of applying FAIR principles to health and genomic information. To do so, we have firstly presented some FAIR initiatives related to health information, like GO FAIR or FAIR4Health. From FAIR4Health, we have taken the steps of the FAIRification workflow. From them, we have identified steps 3 (data de-identification / anonymization) and 6 (license attribution) to be the ones related to security and protection aspects.

In section 3, we have presented the analysis of the different standards associated to de-identification, pseudonymization and anonymization. Moreover, some issues related to license attribution are also introduced. They are further developed in section 4, which describes how MPEG-G [10], an ISO standard to represent genomic information, may provide some of the mechanisms required to solve license attribution issues.

Also related to genomic information, the GA4GH [13] has been working on several recommendations and tools related to security and privacy aspects. One of their produced resources is the Data Use Ontology (DUO) [14], which provides the matching between data use restrictions on genomic data and intended research use requested by researchers. We will study how DUO and other GA4GH specifications may provide some mechanisms to apply FAIR principles to genomic information.

Acknowledgements

This work is partly supported by the Generalitat de Catalunya (2017 SGR 1749). Part of this work is also supported by the FAIR4Health project (Grant agreement 824666, European Commission) through EFMI (European Federation for Medical Informatics).

References

[1] Wilkinson, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[2] FAIR4Health project, https://www.fair4health.eu
[3] GO FAIR, FAIRification process, https://www.go-fair.org/fair-principles/fairification-process
[4] Research Data Alliance, https://www.rd-alliance.org
[5] FORCE11 (the Future of Research Communications and e-Scholarship, https://www.force11.org
[6] FAIRsharing.org, https://fairsharing.org
[7] ISO/IEC 20889:2018, Privacy enhancing data de-identification terminology and classification of techniques, https://www.iso.org/standard/69373.html, https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en
[8] ISO 25237:2017, Health informatics — Pseudonymization, https://www.iso.org/standard/63553.html, https://www.iso.org/obp/ui/ #iso:std:iso:25237:ed-1:v1:en
[9] Creative Commons licenses, https://creativecommons.org/share-your-work/licensing-types-examples/
[10] ISO/IEC 23092 MPEG-G, Genomic Information Representation, 2020. https://www.iso.org/standard/57795.html, https://mpeg-g.org
[11] OASIS, eXtensible Access Control Markup Language (XACML) v3.0, 2017. http://www.oasis-open.org/specs/index.php#xacmlv3.0
[12] Naro, D., PhD Thesis (Advisors: Delgado, J. and Llorente, S.), Security strategies in genomic files, 2020. https://www.tdx.cat/handle/10803/669108
[13] Global Alliance for Genomics and Health (GA4GH), 2018. https://www.ga4gh.org/
[14] GA4GH, Data Use Ontology (DUO), 2019. https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/