Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210810

FAIR Aspects of a Genomic Information Protection and Management System

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

Abstract. To handle genomic information while supporting FAIR principles, we present GIPAMS, a modular architecture. GIPAMS provides security and privacy to manage genomic information by means of several independent services and modules that interact among them in an orchestrated way. The paper analyzes how some security and privacy aspects of the FAIRification process are covered by the GIPAMS platform.

Keywords. Genomic Information, FAIR, Protection, Access control

1. Introduction

Genomic information can be represented using different formats, depending on: kind of information stored (raw, aligned, unaligned, processed in some way, etc.), compressed or not, lossy or lossless, binarized or not, etc. In fact, the format to choose is very much related to the purpose and environment of its use.

In this context, it is also important to take into account that genomic information usually has associated metadata, which can be expressed using XML (eXtensible Markup Language), as in [1],[2], or directly in other formats [3], and can be stored inside [4], or outside the files containing the genomic information they describe (or apply to) [1],[2]. This metadata may describe the information contained in a genomic file, information about the tools or commands used in the processing pipeline, information about what is being studied (medical condition, patient, etc.) or information about security techniques used to protect the genomic information. Furthermore, it might include rules to control the access to the information [4].

Due to the specific characteristics of human genome information, which uniquely identifies a person and her relatives, it is very important to keep it safe, applying security and access control measures. But this may be difficult to achieve when few genomic information is available, since re-identifying data might be a real risk. On the other hand, Findable, Accessible, Interoperable and Reusable (FAIR) principles [5] are a desirable feature for research data, including genomic data.

Once we have the information, then we need the tools to handle it. Again, there are different approaches for this. We have designed a modular architecture, GIPAMS (Genomic Information Protection And Management System), where different

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

functionalities are provided by independent services interacting between them. As indicated in the acronym, security and privacy are key aspects in the design of the platform. Another key aspect is to provide the mentioned FAIR principles. We had previously analyzed security and privacy in the context of FAIR [6] and, in this paper, we use some of those results to validate, from a security point of view, that GIPAMS provides FAIR principles.

In the rest of the paper, we describe our platform architecture and we point out how we provide protection for genomic information while applying FAIR principles.

2. Methods - Platform Architecture

The architecture of the developed platform is an evolution of our original Multimedia Information Protection And Management System, MIPAMS [7]. As it now deals with genomic information, we have called it Genomic Information Protection And Management System, GIPAMS. [8] describes its implementation details. GIPAMS structure is depicted in Figure 1. The different modules are briefly described next:

- User Application: Access point to the whole system. It sends all requests to the Workflow Manager, which redirects to the corresponding module. An access token is required, which is provided by the Authentication Service.
- Workflow Manager: Intermediate module that acts as a unique entry point to the system to facilitate interactions with the other modules and making them transparent to the final user. Before redirecting an operation coming from the User Application, it checks if this operation is authorized using the information inside the access token.
- Authentication Service: User identification server, which uses OAuth 2.0 [9] and JSON Web Tokens [10].
- Genomic Content Service: Deals with genomic archive management, both in reading and writing operations.
- Authorization Service: Validates authorization rules. It mainly receives requests from the Workflow Manager responding to user actions, but other modules may also interact with it.
- Search Service: Performs searches over genomic information.
- Policy Service: Creates authorization rules, which are organized into policies.
- Protection Service: Creates metadata representing protection information associated to genomic information and also applies the defined mechanisms (i.e. encryption, signature, etc.) to the information.
- Report / Track Service: Deals with the reporting of operations done in the system, especially those not authorized. It helps in keeping track of illegal / unusual operations that may indicate an attempt to attack the system.
- Certification Authority: This is not a real module of the system, but something required for its proper functioning.

It is worth noting that this architecture is independent of the kind of information it handles. As indicated, we started with audiovisual information, but we also used it for other types of health information [11].

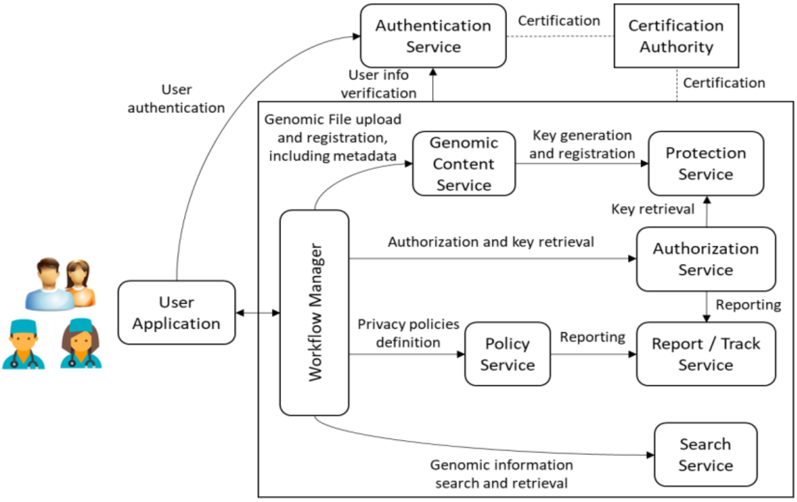

Figure 1. GIPAMS Architecture.

3. Results - FAIR principles in GIPAMS

In order to identify how GIPAMS supports the FAIR principles, we start from the ideas presented in [6], where we analyzed how FAIR principles could be applied to genomic information considering privacy and security. In particular, we discussed in detail steps 3 ("Data de-identification/anonymization") and 6 ("License attribution") of the FAIRification process described in [12]. In the work we are presenting here, however, we focus on step 6, so we consider "license attribution" as the driving concept to analyze how our system GIPAMS is FAIR without losing privacy and security.

Specifically, with respect to license attribution, in [6] we tried to answer to the following four questions: 1) How to express the licenses, 2) How to protect them and guarantee their provenance, 3) How to evaluate their authorization, and 4) How to enforce what they are controlling. The rest of the section provides a first answer to these questions in the context of GIPAMS.

3.1 Expression of licenses

The response to this first question is to use a formal language to facilitate interoperability (the I of FAIR). Rules formally expressed clearly define how to access the information (the A of FAIR). One possibility for the expression of these licenses is to use the eXtensible Access Control Markup Language (XACML) [13] and this is what we have implemented in the Policy Service. XACML allows to express the rules that control who, how and when may access specific genomic information, be it data or metadata. The expression language has an associated mechanism to evaluate the rules, based on standardized requests. In particular, the Policy Service allows the creation of rules, while their evaluation takes place in the Authorization Service.

3.2 Protection of provenance of licenses

The answer to the second question is also implemented in the Policy Service, as the policies and rules created can be protected against modification using XML Signature. Furthermore, this allows to know the origin of the license. This is an extra feature, normally not available in current standards, that we have provided for this service. From a FAIR point of view, this provides support to the A and even to the I, as before, but also, partially, helps in making information Reusable (the R), since we are confident in its origin and lack of modification.

3.3 Authorization upon licenses

As introduced in 3.1, the third question is answered in the Authorization Service, which uses the mechanisms defined in XACML [13]. “XACML Requests” including different attributes like subject, object, action or time conditions have to be defined to check if they fulfill any of the XACML rules stored in the GIPAMS’ Policy Service. The request and the rule are related to an object, which can be any part of the genomic information, including metadata. Again, the A and I from FAIR are supported here.

3.4 Enforcement with licenses

The response to this last question is the GIPAMS platform itself. If the requested action is not authorized, the requested information (which is encrypted for its protection and stored in the Genomic Content Service) will not be provided. It is also possible to keep track of the actions performed in the system by means of the Reporting Module. In this context, it is worth mentioning the Search Service, that would add Findability (F) to the other three FAIR concepts. Although this is not explicit for security, it is however relevant for FAIR.

4. Discussion

We have analyzed a specific aspect of security and FAIR principles (license attribution). Solving the 4 questions raised in section 3 mainly allows to guarantee Accessibility to the genomic information, for the authorized people in the authorized circumstances. Furthermore, GIPAMS also provides Interoperability, since it uses standards for expressing and validating licenses / access rules / policies. Reusability is indirectly provided, as mentioned in 3.2. Finally, Findability is a core part of the platform.

Several GIPAMS’ modules provide this FAIR support, as described in section 3, as Policy, Authorization, Genome Content, Reporting and Search modules. In any case, it is the complete platform who supports the license attribution features.

Another important aspect of a system like GIPAMS is that we might have other “xIPAMS" platforms by providing specific Content services. In other words, our architecture is independent of the kind of content, since we might include different specific Content services. In GIPAMS, we have a Genomic Content service, while in MIPAMS we have a Multimedia Content service. As indicated at the end of section 2, the architecture could be used for other types of health information, as we described in [11]. This means that the provision of security and privacy on one hand, and FAIR principles on the other, is not only valid for genomic information, but also for other eHealth information. This is also very useful when trying to integrate genomic information with current health records.

5. Conclusions

A modular and distributed approach for the management of genomic information facilitates following the FAIR principles. This is accomplished with our Genomic Information Protection And Management System (GIPAMS). We have analyzed how GIPAMS supports the FAIR (Findable, Accessible, Interoperable and Reusable) principles from a security and privacy point of view. We have started from previous work and reached the expected conclusions. Our focus has been on licenses to control the access to information. Some identified GIPAMS' modules, and the complete platform in general, mainly support the Accessibility and Interoperability FAIR principles, but we may also consider the other two.

On the other hand, it is worth noting that, although GIPAMS is intended for the handling of genomic information, other "xIPAMS" platforms may provide services over other eHealth content. Some of our future work concentrates in designing and developing different xIPAMS platforms that would support integration of genomic and other health information, guaranteeing security and privacy and supporting all FAIR principles. We also plan to apply this to a real clinical environment in a new project in Spain.

Acknowledgements

This work is partly supported by the Spanish Government (GenClinLab-Sec, PID2020-114394RB-C31) and by the Generalitat de Catalunya (2017 SGR 1749).

References

[1] National Center for Biotechnology Information (NCBI). 2021. https://www.ncbi.nlm.nih.gov

[2] European Nucleotide Archive (ENA). 2021. https://www.ebi.ac.uk/ena/browser/about

[3] Sequence Alignment / Map (SAM) Format Specification. 2018. https://samtools.github.io/hts-specs

[4] ISO/IEC 23092-3:2020, Genomic information representation - Part 3: Metadata and application programming interfaces (APIs). 2020. https://www.iso.org/standard/75625.html

[5] 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

[6] Delgado J, Llorente S. Security and Privacy when Applying FAIR Principles to Genomic Information. Stud Health Technol Inform. 2020 Nov 23;275:37-41. doi: 10.3233/SHTI200690. PMID: 33227736.

[7] Llorente S, Rodriguez E, Delgado J, Torres-Padrosa V. Standards-based architectures for content management. IEEE MultiMedia. 2012 Nov 29;20(4):62-72.

[8] Delgado, J.; Llorente, S; Reig, G. Implementation of privacy and security for a genomic information system. pHealth 2021. 2021.

[9] IETF. The OAuth 2.0 Authorization Framework, 2012. https://datatracker.ietf.org/doc/html/rfc6749

[10] IETF. JSON Web Token (JWT), 2015. https://datatracker.ietf.org/doc/html/rfc7519

[11] Delgado J, Llorente S. Privacy provision in eHealth using external services. Stud Health Technol Inform. 2015;210:823-7. PMID: 25991269.

[12] GO FAIR, FAIRification process, 2021. https://www.go-fair.org/fair-principles/fairification-process

[13] OASIS, eXtensible Access Control Markup Language (XACML) v3.0, 2017. http://www.oasis-open.org/specs/index.php#xacmlv3.0