A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-561

A Methodology for Fine-Grained Access Control in Exposing Biomedical Data

Alina Trifan[a,1], Johan van der Lei[b], Carlos Díaz[c], José Luís Oliveira[a]
[a] University of Aveiro, IEETA/DETI, Portugal
[b] Erasmus University Medical Center, Rotterdam, Netherlands
[c] Synapse Research Management Partners, Barcelona, Spain

Abstract. Biomedical data integration and processing is a very sensitive issue and a main barrier for research, since it normally implies dealing with private clinical information. To overcome this problem, we propose a solution based on multiple levels of data visibility, combined with a fine-grained access control over the shared data. Through our proposal, on one hand, data custodians can decide the level of detail at which they want to share data, in a flexible manner that can be adjusted along the time. On the other hand, adequate permissions are provided to the users that want to access the data, according to their role and research plan.

Keywords. biomedical data exposure, data privacy, access control, data sharing

1. Introduction

In recent years, several solutions for biomedical data integrations were proposed [1–3]. One of such initiatives is the European Medical Information Framework[2] (EMIF), which aims at leveraging the access to patient data, by aggregating biomedical data sources otherwise found in disparate locations and systems. The main goal of these initiatives is to encourage collaborative research and clinical data reuse. Clinical studies are often a burden for researchers, who first need to identify data sources, request access to the data, access each data source, many times in an isolated way, and then conduct the study. Despite enabling collaborative research and clinical data reuse, data integration has to take into account privacy and confidentiality issues and ensure that no private data are revealed. We propose a solution for the exposure of private biomedical data in a safe way, in which the data custodian controls, at the core of the system, to whom and how much of data can be shared.

2. Background

Health information is regarded by many as the most confidential of all types of personal information [4]. Protecting these data is crucial and it is part of one of the three security goals that have to be met when exposing it: confidentiality, integrity and availability [5].

[1] Corresponding Author: alina.trifan@ua.pt

[2] http://www.emif.eu

To facilitate healthcare evolution, patient data usually need to be widely disseminated [6]. This, however, cannot imply the disclosure of private and sensitive information.

Along with the proliferation of digital biomedical data, the concern of maintaining data privacy, while still making use of the benefits of data exposure and integration, has arose [7]. Most often, researchers rely on data anonymization or de-identification in order to protect biomedical data privacy [8]. However, the detail and diversity of information collected in the context of healthcare and biomedical research is increasing at an unprecedented rate and the ready availability of such large volumes of detailed data has also been accompanied by privacy concerns [9].

Several projects have been developed for exposing and discovering de-identified health data in efficient ways [10–12]. Most of these solutions represent isolated software tools or projects designed to support specific research, with a clear purpose, but new solutions for redefining the way data privacy is handled need to be addressed. In this paper we provide a different perspective on how data privacy can be maintained when biomedical data is exposed.

3. Method

The main purpose of our work was to facilitate biomedical data exposure and sharing, but at the same time, to ensure fine-grained access control to these biomedical data. For that, we developed a methodology that tackle data privacy from a different perspective, in which real data sources are characterized based on a schema that can be easily defined by any data custodian. The data custodian controls right upfront how into detail s/he wants to expose a data source. Several levels of information can be exposed, from summarized views to raw data, accessible in a controlled and possibly remote environment. Moreover, a role-based access control (RBAC) assures that an access policy can be tailored to combine the access constrains with the needs of biomedical researchers. By conjugating these two orthogonal perspectives, we designed an approach for exposing biomedical data sources, at different levels of details, while preserving data privacy.

3.1. Data exposure

Data exposure is provided from a drill-down perspective, where at least three different views can be explored: a general one, over metadata extracted through the characterization schema, a summarized view of aggregated information and a deep view, that can extend up to the exposure of raw data, under controlled conditions (Figure 1). Through a dynamic characterization schema, data custodians are given the opportunity to ex- pose their data on a multi-level visibility approach. The schema architecture is flexible enough to allow the data custodian to expose as much or as little information that s/he understands necessary and can be extended or diminished at any time. Its definition is a straightforward operation that can be done by any data custodian by following some simple formatting rules.

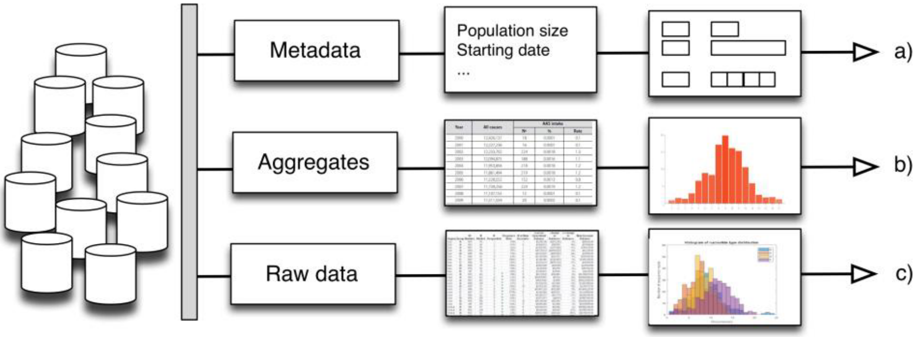


Figure 1. A multi-level model for clinical data exposure. The example includes three distinct paths: a) a general view over metadada; b) an aggregate view, over summary-level data; c) a deeper view, over raw data.

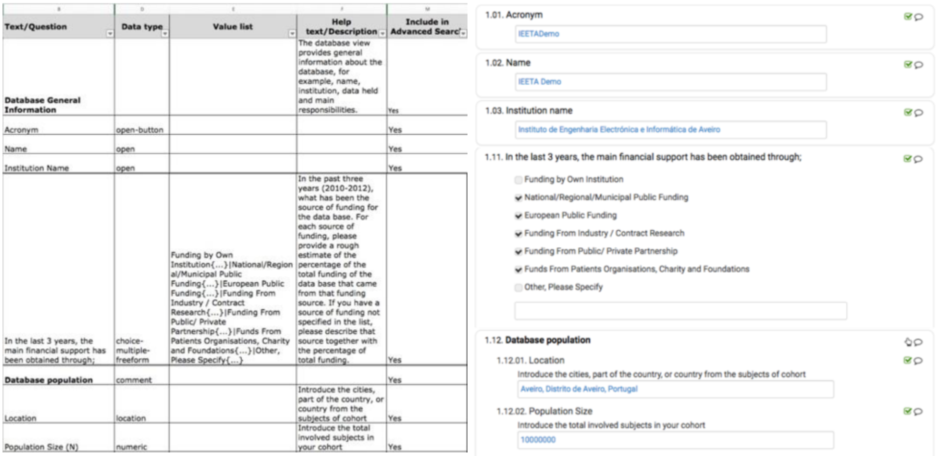
3.2. Data access control

Once the data is published, registered users can access it, according to their research needs and taking into consideration the permissions they are granted. A user permission system further refines the multiple layers of data access control.

Different user profiles and roles are contemplated. Combined, they provide the right amount of access to the right amount of data. According to the user profile, this approach also allows managing the access to distinct functionalities and tools that allow exploring the published data. Moreover, users can be organized in groups, simplifying the management of profiles, roles and access permission. User groups can be created and granted different data or functionality permissions at any time. A rule of thumb is that default users can browse, search and compare public metadata exposed by data custodians. Data custodians, for example, are specialized users that have been granted the permission to add information about a data source. More advanced users can be granted permissions for evaluating requests to a deeper data view and to manage the proper tools for analyzing and eventually granting the requested access.

4. Results and Discussion

To address the technical requirements raised by the proposed methodology, we developed a solution for data characterization with multiple levels of exposure. The schema for metadata extraction can be constructed as simple as a spreadsheet by the data custodian, who is provided with a set of formatting options. The schema is then rendered in a user-friendly way (Figure 2), where different views are available for the end-user. These views are in practice, representations of data according to visibility levels that were granted. A general view displays the metadata that best characterizes a given data source. The summarized view, for example, presents population characteristics, or average measurements, while the raw view exposes real data if research agreements and adequate protocols are in place.



Text/Question	Data type	Value list	Help text/Description	Include in Advanced Searc'
Database General Information			The database view provides general information about the database, for example, name, institution, data held and main responsibilities.	Yes
Acronym	open-button			Yes
Name	open			Yes
Institution Name	open			Yes
In the last 3 years, the main financial support has been obtained through;	choice-multiple-freeform	Funding by Own Institution{...}|National/Regional/Municipal Public Funding{...}|European Public Funding{...}|Funding From Industry / Contract Research{...}|Funding From Public/ Private Partnership{...}|Funds From Patients Organisations, Charity and Foundations{...}|Other, Please Specify{...}	In the past three years (2010-2012), what has been the source of funding for the data base. For each source of funding, please provide a rough estimate of the percentage of the total funding of the data base that came from that funding source. If you have a source of funding not specified in the list, please describe that source together with the percentage of total funding.	Yes
Database population	comment			Yes
Location	location		Introduce the cities, part of the country, or country from the subjects of cohort	Yes
Population Size (N)	numeric		Introduce the total involved subjects in your cohort	Yes

1.01. Acronym

IEETADemo

1.02. Name

IEETA Demo

1.03. Institution name

Instituto de Engenharia Electrónica e Informática de Aveiro

1.11. In the last 3 years, the main financial support has been obtained through;

- □ Funding by Own Institution
- ☑ National/Regional/Municipal Public Funding
- ☑ European Public Funding
- ☑ Funding From Industry / Contract Research
- ☑ Funding From Public/ Private Partnership
- ☑ Funds From Patients Organisations, Charity and Foundations
- □ Other, Please Specify

1.12. **Database population**

1.12.01. Location

Introduce the cities, part of the country, or country from the subjects of cohort

Aveiro, Distrito de Aveiro, Portugal

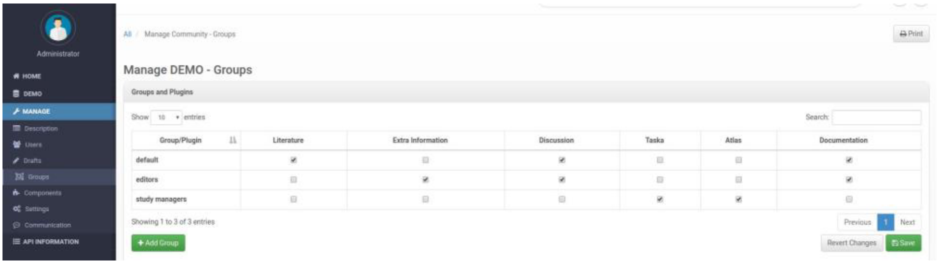
1.12.02. Population Size

Introduce the total involved subjects in your cohort

10000000

Figure 2. Mapping from the schema definition into HTML5 forms. On the left, a view of some fields from the schema template (open-text, multiple choice, numeric and location - these are just a few of the formatting options that a data custodian can use). On the right, their rendering on the general view.

Once the several data views are made public by the data custodian, granular access control can be given to users, according to their research profiles. Such access refers to tools that they can use along with the exposed metadata, permissions to export the metadata or the result of comparative search among different data sources. In Figure 3 we present an example in which three user groups are defined: default, editors and study managers. Each of these groups can access different functionalities, designed to meet their research needs and authorization agreements. In this example, we include two components that can be used in a research study, TASKA[3] and ATLAS[4], in order to illustrate our approach for granting distinct permissions to specialized users. TASKA is a work- flow system intended for the management of research studies, while ATLAS provides a unified interface to patient level data and analytics.



Manage DEMO - Groups

Groups and Plugins

Show 10 entries

Search:

Group/Plugin	Literature	Extra Information	Discussion	Taska	Atlas	Documentation
default	☑	□	☑	□	□	☑
editors	□	☑	☑	□	□	☑
study managers	□	□	□	☑	☑	□

Showing 1 to 3 of 3 entries

Figure 3. User groups and permissions. Each column represents a different functionality, that can be used either at the level of the data source or at a higher level of data visibility, such as the aggregated view.

[3] https://bioinformatics.ua.pt/taska

[4] http://www.ohdsi.org/web/atlas/

5. Results and Discussion

In this paper, we proposed a two-fold approach for promoting biomedical data exposure and discovery, while addressing privacy concerns. The first strand of our solution provides the support for a data custodian to control the visibility of his/her data. The second one further refines data access control through user permission policies. The methodology for fine-grained access that we designed is already being utilized by several research communities, showing that biomedical data exposure and collaboration is possible, without breaking any data privacy.

Acknowledgements

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n. 115372).

References

[1] Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. Research Electronic Data Capture (REDCap)-a Metadata-driven Methodology and Workflow Process for Providing Translational Research Informatics Support. Journal of Biomedical Informatics, 42(2):377–381, April 2009.

[2] Sharon F Terry. The Global Alliance for Genomics & Health. Genetic Testing and Molecular Biomarkers, 18(6):375–376, 2014.

[3] Chao Pang, David van Enckevort, Mark de Haan, Fleur Kelpin, Jonathan Jetten, Dennis Hendriksen, Tommy de Boer, Bart Charbon, Erwin Winder, K Joeri van der Velde, et al. Molgenis/connect: a System for Semi-Automatic Integration of Heterogeneous Phenotype Data with Applications in Biobanks. Bioinformatics, 32(14):2176–2183, 2016.

[4] José Luis Fernández-Alemán, Inmaculada Carrión Senõr, Pedro Ángel Oliver Lozoya, and Ambrosio Toval. Security and privacy in electronic health records: A systematic literature review. Journal of biomedical informatics, 46(3):541–562, 2013.

[5] Sebastian Haas, Sven Wohlgemuth, Isao Echizen, Noboru Sonehara, and Gunter Muller. Aspects of privacy for electronic health records. International journal of medical informatics, 80(2):e26–e31, 2011.

[6] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jian Sun. Toward smarter healthcare: Anonymizing medical data to support research studies. IBM Journal of Research and Development, 58(1):9–1, 2014.

[7] Fengjun Li, Xukai Zou, Peng Liu, and Jake Y Chen. New threats to health data privacy. BMC bioinformatics, 12(12):S7, 2011.

[8] Noman Mohammed, Benjamin Fung, Patrick CK Hung, and Cheuk-Kwong Lee. Centralized and distributed anonymization for high-dimensional healthcare data. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(4):18, 2010.

[9] Bradley A Malin, Khaled El Emam, and Christine M O'keefe. Biomedical data privacy: problems, perspectives, and recent advances, 2013.

[10] Xiaoling Chen, Anupama E Gururaj, Burak Ozyurt, Ruiling Liu, Ergin Soysal, Trevor Cohen, Firat Tiryaki, Yueling Li, Nansu Zong, Min Jiang, et al. Datamed–an open source discovery index for finding biomedical datasets. Journal of the American Medical Informatics Association, 2018.

[11] James J Cimino, Elaine J Ayres, Lyubov Remennik, Sachi Rath, Robert Freedman, Andrea Beri, Yang Chen, and Vojtech Huser. The National Institutes of Healths biomedical translational research information system (btris): design, contents, functionality and experience to date. Journal of biomedical informatics, 52:11–27, 2014.

[12] Lucia Monaco, Marco Crimi, and Chiuhui Mary Wang. The challenge for a European network of biobanks for rare diseases taken up by RD-Connect. Pathobiology, 81(5-6):231–236, 2014.