Published online: 2020-07-03

From Raw Data to FAIR Data: The FAIRification Workflow for Health Research

A. Anil Sinaci[1] Francisco J. Núñez-Benjumea[2] Mert Gencturk[1] Malte-Levin Jauer[3] Thomas Deserno[3] Catherine Chronaki[4] Giorgio Cangioli[4] Carlos Cavero-Barca[5] Juan M. Rodríguez-Pérez[5] Manuel M. Pérez-Pérez[5] Gokce B. Laleci Erturkmen[1] Tony Hernández-Pérez[6] Eva Méndez-Rodríguez[6] Carlos L. Parra-Calderón[2]

[1] SRDC Software Research Development and Consultancy Corporation, Ankara, Turkey
[2] Group of Research and Innovation in Biomedical Informatics, Biomedical Engineering and Health Economy, Institute of Biomedicine of Seville/Virgen del Rocío University Hospital/CSIC/University of Seville, Seville, Spain
[3] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany
[4] Health Level Seven International Foundation, Brussels, Belgium
[5] Atos, Group of Health, Atos Research and Innovation (ARI), Madrid, Spain
[6] Department of Library and Information Sciences, Universidad Carlos III de Madrid, Madrid, Spain

Address for correspondence A. Anil Sinaci, PhD, SRDC Software Research Development and Consultancy Corporation, ODTU Teknokent Silikon Bina K1-16 06800 Cankaya, Ankara, Turkey (e-mail: anil@srdc.com.tr).

Methods Inf Med 2020;59:e21–e32.

Abstract

Background FAIR (findability, accessibility, interoperability, and reusability) guiding principles seek the reuse of data and other digital research input, output, and objects (algorithms, tools, and workflows that led to that data) making them findable, accessible, interoperable, and reusable. GO FAIR - a bottom-up, stakeholder driven and self-governed initiative - defined a seven-step FAIRification process focusing on data, but also indicating the required work for metadata. This FAIRification process aims at addressing the translation of raw datasets into FAIR datasets in a general way, without considering specific requirements and challenges that may arise when dealing with some particular types of data.

Objectives This scientific contribution addresses the architecture design of an open technological solution built upon the FAIRification process proposed by "GO FAIR" which addresses the identified gaps that such process has when dealing with health datasets.

Methods A common FAIRification workflow was developed by applying restrictions on existing steps and introducing new steps for specific requirements of health data. These requirements have been elicited after analyzing the FAIRification workflow from different perspectives: technical barriers, ethical implications, and legal framework. This analysis identified gaps when applying the FAIRification process proposed by GO FAIR to health research data management in terms of data curation, validation, deidentification, versioning, and indexing.

Results A technological architecture based on the use of Health Level Seven International (HL7) FHIR (fast health care interoperability resources) resources is proposed to support the revised FAIRification workflow.

Keywords
- interoperability
- data science
- data curation
- data anonymization
- metadata

received
July 31, 2019
accepted after revision
May 6, 2020

DOI https://doi.org/10.1055/s-0040-1713684.
ISSN 0026-1270.

Discussion Research funding agencies all over the world increasingly demand the application of the FAIR guiding principles to health research output. Existing tools do not fully address the identified needs for health data management. Therefore, researchers may benefit in the coming years from a common framework that supports the proposed FAIRification workflow applied to health datasets.

Conclusion Routine health care datasets or data resulting from health research can be FAIRified, shared and reused within the health research community following the proposed FAIRification workflow and implementing technical architecture.

Introduction

The FAIR (findability, accessibility, interoperability, and reusability) data principles were first published in 2016.[1] FAIR seeks the reuse of data and other digital research output and objects (algorithms, tools, and workflows that led to that data) making them findable, accessible, interoperable, and reusable. These principles consider applications and computational agents as stakeholders with the capacity to find, access, interoperate, and reuse data with none or minimal human intervention. They also recognize the importance of an automated process for computational support to deal with intensive data processes. As stated by Mons et al,[2] FAIR refers to a set of principles, focused on ensuring that research objects are reusable, are actually reused, and in this way become as valuable as is possible. They deliberately do not specify technical requirements but deliver a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of creators and users.

The first draft of the FAIR data principles was born in January 2014 at the Lorentz Center in Leiden, the Netherlands, by a community of scholars, librarians, archivists, publishers, and research funders as a part of the Future of Research and Communications and e-Scholarship group (FORCE11, *https://www.force11.org/*). As early as July 2016, the European Union (EU) published the Guidelines on FAIR data management in Horizon 2020.[3] The principles are also explicitly mentioned in the new open data and reusable public sector information (PSI) directive,[4] and the European Open Science Cloud (EOSC) focuses on enabling FAIR data and principles. In the United States, the National Institutes of Health (NIH) also support the FAIR principles[5] and it can be said that the most important research funding agencies and international organizations support or have adopted these principles.

Among ongoing initiatives addressing the application of FAIR data principles in practice, GO FAIR is a prominent one. GO FAIR is a bottom-up, stakeholder-driven, and self-governed initiative that aims to implement the FAIR data principles. It offers an open and inclusive ecosystem for individuals, institutions and organizations working together through implementation networks (INs).[6] The FAIR data principles apply not only to data, but also to metadata, supporting infrastructure (e.g., search engines) and other research outputs. At the metadata level, findability and accessibility requirements must be addressed, while interoperability and reuse require more efforts at the data level. GO FAIR defined a seven-step FAIRification process[7] focusing on data, but also indicating the required work for metadata alignment. The FAIRification process was conceived as a set of step-by-step operations that should be performed over data and related metadata to achieve its FAIRness. According to GO FAIR, the steps involved in this process are as follows: (1) retrieve non-FAIR data, (2) analyze the retrieved data, (3) define the semantic model, (4) make data linkable, (5) assign license, (6) define metadata for the dataset, and (7) deploy/publish FAIR data resource.

This FAIRification process aims at addressing the translation of raw datasets into FAIR datasets in a general way, without considering specific requirements, and challenges that may arise when dealing with some particular types of data, such as health data. Beyond the technical interoperability challenges of connecting various information systems and using analytical methodologies able to cope with the growing amount of data, managing and reusing health data also poses significant challenges from ethical, legal, and privacy points of view.[8–10] Furthermore, these issues are usually interwoven and must be tackled with a common strategy.

Ethical considerations for health research commonly rely on the World Medical Association (WMA) Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects (revised 2014).[11] Recently, the WMA released the Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks[12] to address ethical issues derived from reusing health data beyond its original purpose (i.e., for research and innovation, mainly). To comply with this declaration when reusing health data, reasonable efforts to obtain voluntary and informed consent must be sought, and dignity, autonomy, privacy, and confidentiality of patients must be protected among other considerations. To cope with these ethical principles, national regulations have been developed in the last years. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) first released in 1996 and then amended in 2009 through the Health Information Technology for Economic and Clinical Health (HITECH) Act,[13] provides the legal framework for individuals and health professionals about health information privacy. HIPAA contemplated that most research would be conducted

by universities and health systems, but today much of the demand for data emanates from private companies at which ethical review boards (ERBs) and privacy boards may be weaker or nonexistent.[14] Moreover, the increasing availability of data generated outside health care settings, coupled with advances in computing, undermines the historical assumption that data can be forever deidentified.[15] In the EU, the General Data Protection Regulation (GDPR)[16] entered into force in 2018 and prevails over potentially clashing national regulations. GDPR sets forth a set of principles that must be followed when processing personal information: (1) lawfulness, fairness, and transparency; (2) purpose limitation; (3) data minimization; (4) accuracy; (5) storage limitation; and (6) integrity and confidentiality. Besides, it enforces compliance with accountability by encouraging good practices, such as implementing data protection policies and security mechanisms, documenting any processing activity, and carrying out data protection impact assessments, among others. Nonetheless, if at some point, it is possible to achieve total anonymization that would guarantee the absolute impossibility of reidentifying the data subject, anonymized data would cease to have the status of personal data under the GDPR perspective. It would therefore be possible to process such data without having to comply with the data protection requirements.

However, patient privacy is not the only challenge when processing health data for a secondary use. Extracting structured and accurate information from unstructured reports, such as narrative sections in electronic health records (EHRs), is a common challenge when performing research over big cohorts of subjects within the Natural Language Processing (NLP) domain.[17] However, NLP performance highly depends on the study objective, clinical domain and language,[18–20] and therefore, although promising, it is still far from the general application in clinical settings.

The FAIRification process[7] supported by GO FAIR was developed considering data management needs derived from research outputs to optimize their reusability. Health research datasets are commonly static in the sense that they represent the status of a sample or patient at specific time points, usually before and after the application of an intervention, to observe its impact on the sample or patient. This argument is not valid when researchers want to reuse health data gathered for other purposes, such as routine care, since this type of data are not gathered during scheduled consultations but under the continuum of health care delivery. How to appropriately manage the update and versioning of these datasets in a research environment is still an open issue.[21–23]

Objectives

The overall objective of this work is to encourage the health research community to FAIRify, share and reuse their datasets derived from publicly funded activities (both research and health care) by facilitating the FAIRification process, and demonstrating the potential impact of such strategy on health outcomes and health research. In line with this objective, an intuitive, user-centered technological solution is being developed to enable the transformation of raw datasets into FAIR datasets. These datasets can be gathered from different sources, such as specific collections of health research data, EHRs, personal health records (PHRs), as well as other datasets addressing health, social, and environmental determinants, among others.

In this context, this paper presents the architecture design of an open technological solution built upon the FAIRification process proposed by GO FAIR closing the gaps of this process for health datasets, thus providing the health research community with a common, standards-based, legally-compliant FAIRification workflow for health data management. The actual implementation of the proposed architecture has been initiated as an open-source activity (https://github.com/fair4health), developing a set of software tools addressing different steps of the FAIRification workflow.

Methods

This work adapts the FAIRification process introduced by GO FAIR to health data and proposes a common FAIRification workflow by applying restrictions on existing steps and introducing new steps for specific requirements of health data. These requirements have been elicited after analyzing the needs that such a FAIRification workflow may have from different perspectives: technical barriers, ethical implications, and legal framework. The analysis of technical barriers was based on the conclusions extracted by a focus group of experts after discussing about the conclusions reached by Wilkinson et al[24] and the "turning FAIR into reality" high level expert group report.[25] The analysis of ethical implications follows the conclusions of a focus group that reviewed the results of an open survey.[26–30] Finally, a comprehensive review of the EU GDPR and national legislations in Spain, Italy, Switzerland, and Serbia, in the context of FAIR4Health project,[31] was performed. The revision included the following regulations:

- EU: regulation 2016/679 of the European Parliament and of the Council of April 27, 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC.
- Italy: legislative decree no. 196/2003, containing the "code for the protection of personal data"; General authorization no. 9/2016 to process personal data for scientific research purposes; general authorizations no. 1/2016, 3/2016, 6/2016, 8/2016, and 9/2016 that are compatible with the regulation and legislative decree no. 101/2018; deontological rules for statistical or scientific research treatments, no. 515/2018.
- Serbia: law 87/2018–54 on personal data protection.
- Spain: organic law 3/2018, of 5 December, on the protection of personal data and the guarantee of digital rights; law 14/2007, of 3 July, on biomedical research; law 33/2011, of 4 October, general public health.
- Switzerland: data protection framework in CH; Federal Act data protection; Federal Data Protection and Information Commissioner (FDPIC); the Human Research Act (HRA).

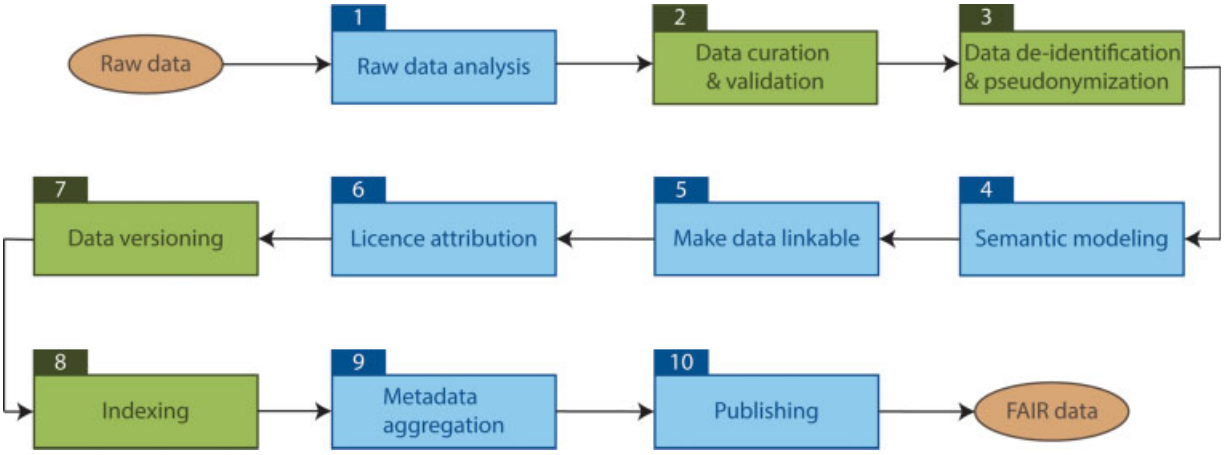

Fig. 1 The FAIRification workflow for health data. The blue boxes come from the GO FAIR (an initiative implementing FAIR data principles) process, while the green boxes are newly introduced steps to meet the specific challenges of health data. FAIR, findability, accessibility, interoperability, and reusability.

- As a result of the comprehensive analysis performed, this paper introduces the proposed FAIRification workflow tailored to the specific needs and requirements posed by the use of health data as shown in ►**Fig. 1**. Blue boxes (1,4,5,6,9,10) come from the GO FAIR process and this paper proposes restricted actions for those steps, while the green boxes (2,3,7,8) are newly introduced steps to meet the specific challenges of health data, such as health data curation bearing the clinical concepts in mind, and highly sensitive nature of health data.

Raw Data Analysis

The raw data analysis inspects the content of the data to find out which concepts are represented, what is the structure within and among the data element concepts, and which is the storage and serialization format of the data elements. At this step, analyzing health data requires extra intervention first to take the clinical data models into account. In this domain, there are several standardization efforts for health data management, for example, research data to be repurposed for future studies are expected to conform to clinical data standards. Hence, this step should be aware of the well-established and widely used standards in clinical care and clinical research such as Health Level Seven International (HL7) clinical document architecture (CDA),[32] HL7 fast health care interoperability resources (FHIRs),[33] Open EHR,[34] and observational medical outcomes partnership (OMOP) common data model (CDM).[35] In addition to data modeling, health data utilizes several coding schemes and terminology systems which needs to be considered during the raw data analysis step.

Data Curation and Validation

This step is directed toward increasing the quality of the dataset for research purposes. This step may be of little help for raw datasets coming from research, which are usually well curated. However, health data extracted from routine care usually need to be repurposed and validated to be useful for research. During data curation, data fields, types, and values (metadata) are characterized, and clinical concepts, such as data elements or fields for diagnostics, medications, laboratory results, etc., are extracted. The curated data should be validated against known quantitative relationships and expected values and should conform to semantic model (step 4) defined for the FAIRification workflow, that is, the predefined target data model through a set of structural rules. Moreover, the data itself (i.e., the value of a systolic blood pressure measurement) should conform to the semantic rules exposed by the data element or attribute itself.

Data Deidentification and Pseudonymization

Once the dataset is curated, validated, and has relevant metadata aggregated, the next step would be to deidentify and/or pseudonymize the dataset to enable its sharing without comprising the data subjects' rights regarding privacy issues (Cf. 3.4.2, GDPR). The decision to apply deidentification and/or pseudonymization to the dataset will depend on the purpose for which the dataset has been developed. For instance, in case there is a need to update the dataset from time to time as more data from the same data subjects are available, pseudonymization techniques, such as one-way encryption of identifiers, could be applied. Specific registries could be located and updated with new information without disclosing sensitive information in the process.

Deidentification techniques could be developed based on the HIPAA deidentification Standard of Protected Health Information[36] that identifies data types that should be dropped from a health dataset to minimize the risk of reidentification of data subjects.

Pseudonymization techniques can be based on replacing personal identifiers with artificial identifiers or pseudonyms. This is usually performed over data subjects' unique identifiers, such as national/passport ID number and health ID number This pseudonym along with the rest of personal identifiers (full name, telephone number, address, etc.) is stored in a separate file. To this end, there are multiple one-way encryption algorithms that could be applied, such as the Secure Hash Algorithm (SHA-x).[37,38] It should be noted that, although nowadays one-way encryption algorithms are considered secure, they all are

exposed to brute force attacks which will become more feasible as information technologies evolve and greater processing capacity is made available to the general public.

Apart from deidentification of directly identifying attributes, such as patient IDs and names, deidentification of other elements of the datasets that may act as quasi-identifiers, such as dates (such as birth, death, admission, discharge, visit, and specimen collection), locations (such as postal codes, hospital names, and regions), race and ethnicity, and in some cases rare diagnoses, should be addressed as well. Different methods for deidentification, such as fuzzing (adding "noise" to an atomic data element), generalization (making an atomic data element less specific and longitudinal consistency (modifying data so that it is shifted by a specific amount), can be considered for these quasi-identifiers.[39]

Semantic Modeling

This step involves defining a "semantic model" for the dataset, which describes the meaning of entities and relations in the dataset accurately, unambiguously, and in a computer-actionable way. Depending on the dataset, defining a proper semantic model may require a significant effort, even for experienced data modelers. A good semantic model should represent a consensus view in a particular domain, for a particular purpose. Therefore, it is good practice to comply with existing models that are resulting from standardization efforts. Data curation should map the raw data conforming to such a standard-based data model. In the health domain, in addition to data model standards like HL7 CDA,[32] HL7 FHIR,[33] or OMOP CDM,[35] there are widely used vocabularies for health-related terms. The semantic model for the dataset should incorporate vocabularies, terminology systems, coding standards such as International Classification of Disease (ICD),[40] Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), and [41] Logical Observation Identifiers Names and Codes (LOINC).[42].

Make Data Linkable

Raw data can be transformed into linkable data by applying the semantic model defined in the previous step. Currently, this is done using semantic web and linked data technologies. This step promotes interoperability and reuse, facilitating the integration of the data with other types of data and systems. However, the user should evaluate the feasibility of this step for the given dataset. It is a sensible thing to do for many types of data (e.g., structured data), but it may not be relevant for other types (e.g., the pixels in images).

License Attribution

The use of licensing attributions applied to health datasets must be subject to the regulatory framework in force for each data owner, especially when it comes to sensitive data. The importance of setting clear license terms is required for reusability of the dataset. Therefore, the license attribution for the dataset should be stated clearly, as well as the process by which an external requester could request permission to reuse the dataset.

Data Versioning

The Research Data Alliance (RDA) Working Group on data citation released a set of 14 recommendations regarding data versioning[43] to enable precise identification of every subset and version of data used, supporting reproducibility of processes, sharing and reuse of data. Data versioning should be handled following the international standards, best practices, and recommendations such as those of RDA. The standard recommended for data versioning is the Reference Model for an Open Archival Information System (OAIS), ISO 14721:2003.[44] The reference model embraces six entities and three information packages, which have been used previously toward the design of a repository for standardized medical image and signal case data annotated with ground truth[45]:

1. Ingest: the first stage of OAIS includes the receipt of submission information package (SIP) that is created and submitted by the producer (user or system), and recording of data provenance in archival information package (AIP), which is sent to the persistence archive.
2. Archival storage: in this step, semantic enrichment and linking of data are handled and fed back into the AIPs.
3. Data management: the OAIS model has a component for administering the database and performing queries. This component is enhanced with data privacy mechanisms and deidentification procedures, applicable when health data are processed. Altogether, ingest, archival storage, and data management make sure that FAIR data are generated.
4. Preservation planning: this OAIS component is designed to ensure long-term accessibility of data. The persistent archive holds a history change log, particularly including the versioning of data, and all information required to recreate any dissemination information package (DIP) that has been created previously.
5. Access: as the data and metadata such as annotations are subject to modifications within their full lifetime span, the user can loop a modified DIP back as SIP, where the next ingest step includes a versioning of data. This step ensures reproducibility on any experimental dataset that has been delivered in the past.
6. Administration: the last OAIS component is related to the overall operation of FAIR data and includes, for instance, agreements with data producers and consumers.

Indexing

Indexing is an important step for the health data since searching over these datasets is one of the ultimate goals of FAIRification. Especially for the secondary use of health data, executing eligibility queries over a population, or querying individual records of a single patient is crucial. Each versioned data needs to be indexed with respect to the possible search parameters over the records. HL7 FHIR provides an inherent functionality for configuring the search parameters on the resource types and this step utilizes this capability of FHIR with the design of data versioning.

Metadata Aggregation

This action is performed to state the dataset data provenance, increase its quality, and understandability, thus enabling its findability and reusability in further research studies. There are many metadata standards and vocabularies already available to the scientific community. However, a metadata schema for clinical research in general based on the DataCite standard has been published by researchers from the European Clinical Research Information Network (ECRIN) group addressing the general needs that health research projects have in common.[46] This schema should serve as a starting point for those public health care providers and health researchers wishing to FAIRify their datasets for research purposes.

Publishing

Data publishing is the process of making FAIR health datasets available to a separate storage device for long-term retention/preservation. For health datasets, this is not a trivial issue, since data types and sizes may hugely vary depending on the original sources. Publishing datasets in an external repository does not imply making the data open, given that some repositories make datasets available under licenses similar to the "on demand, upon approval."

Results

The FAIRification workflow presented in this paper is designed by focusing on the specific needs and requirements of health data. Any system aiming to transform raw health data into FAIR health data should provide effective solutions, especially for the newly introduced steps in the workflow. In this regard, in this section, we present an architecture (►Fig. 2) that meets the challenges exposed by our newly introduced steps within the FAIRification workflow. The remaining ones (GO FAIR steps) can be implemented through organization-wide business processes and adopting health-specific software frameworks.

As depicted in ►Fig. 2, an FHIR Repository resides at the core of the architecture as the health data repository. FHIR[33] is a next generation standards framework created by HL7. FHIR application programming interfaces (APIs) are built from a set of modular components called "Resources." Plugging into FHIR APIs, these resources can easily be assembled into working systems that solve real-world clinical and administrative problems at a fraction of the price of existing alternatives. Utilizing the HL7 FHIR standard within the data source facilities provides support and is an enabling factor for data FAIRification in many aspects as follows:

- HL7 FHIR assigns a globally unique and persistent logical identifier (i.e., Object Identifier) to each resource (findable).
- The location of a resource instance is an absolute Uniform Resource Identifier (URI) constructed from the server base address at which the instance is found, the resource type and the logical ID, such as *http://test.hl7.fhir.org/rest/Patient/123* (where 123 is the logical Id of a patient resource). When the location is an Hypertext Transfer Protocol (HTTP) address, this address can generally be used to retrieve or manipulate the resource (accessible).
- HL7 FHIR provides a "formal, accessible, shared, and broadly applicable" way to represent (health) information (interoperable).
- Each HL7 FHIR resource includes a rich set of attributes to describe the most relevant data and metadata; a formal extension mechanism is also specified by that standard to cover additional requirements (reusable).

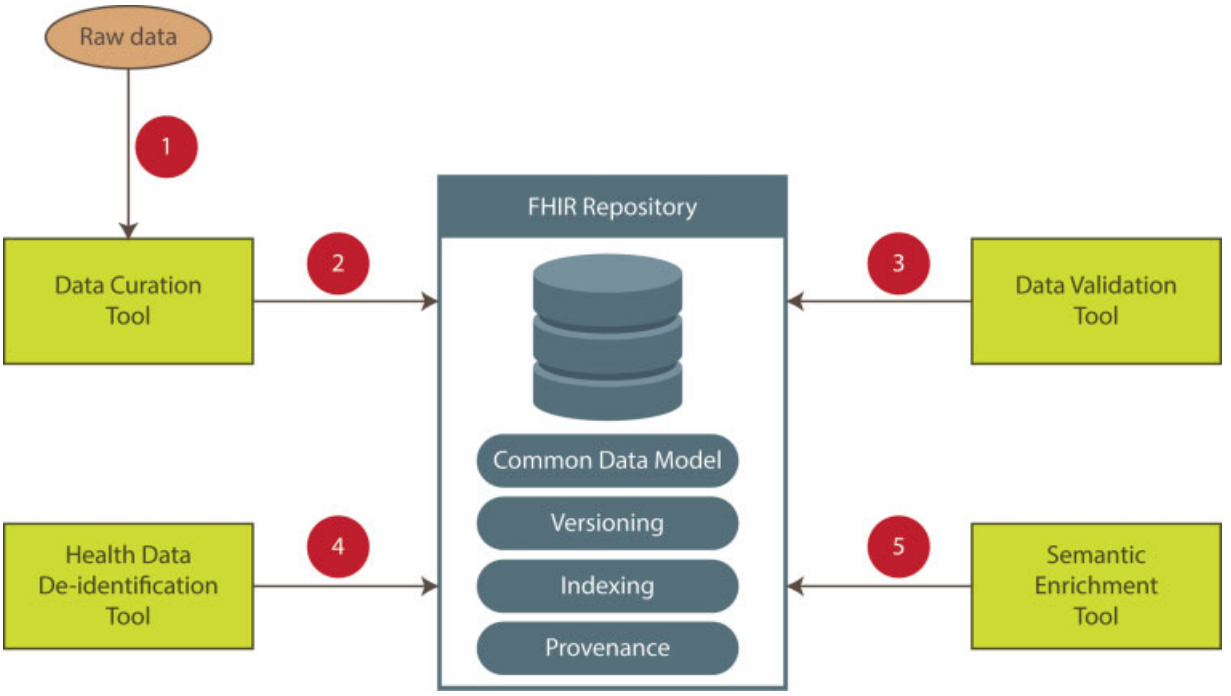

Fig. 2 An architecture implementing the FAIRification workflow for health data. The numbered red circles refer to FAIRification workflow steps that are explained in the text. FAIR, findability, accessibility, interoperability, and reusability; FHIR, fast health care interoperability resources.

- FHIR enables to define a Semantic Model through FHIR profiles which are a set of constraints on a resource represented as a structure definition.
- Data versioning is an inherent functionality that ships with FHIR. Data provenance is implemented with the FHIR provenance resources.

Around the FHIR Repository, there exist several components for transforming the raw data into FAIR data. These components are expected to operate in the same order as the FAIRification workflow. Detailed information about these components are provided in the following subsections.

Data Curation Tool and Data Validation Tool

According to the FAIRification workflow for health data, after raw data analysis, data curation and validation are performed first. The aim of Data Curation and Validation tool is to not only increases the quality of data for research purposes but also made the data accessible through a standard API, that is, HL7 FHIR, so that interoperability of FAIR data can be satisfied.

Data curation and validation of health data in raw format can be performed in several steps that are illustrated in ►Fig. 3.

1. The data manager connects to the data source and with the help of data source analyzer looks at the metadata in the data source. Various data sources with different data formats can be connected including comma-separated values (CSV) files and relational databases with custom information models, as well as standard interfaces, such as picture archiving and communication system (PACS) and integrating the healthcare enterprise (IHE) cross-enterprise document sharing (XDS) profiles[47] which provide digital imaging and communications in medicine (DICOM)[48] and the Consolidated CDA (C-CDA)[49] formatted medical data respectively.
2. The Data Manager maps the source data elements (the metadata, i.e., patient date-of-birth field) to the target CDM (the predefined FHIR profile) through Metadata Mapper.
3. Since the CDM is described as a FHIR profile, Metadata Mapper communicates with the FHIR Repository to show the target data elements (the resource types and attributes) to the Data Manager associated with the FHIR repository.
4. Once the mappings are set, data are transferred to Data Transformer.
5. Data Transformer accesses the data source and retrieves the original data.
6. Data Transformer transforms the data to the FHIR Repository, conforming to the CDM, based on the given mappings.
7. Data Validator is an inherent part of the FHIR Repository. During the transformation, it validates each transaction in terms of conformance to the defined FHIR profile. Due to this validator, it is not possible to insert resource instances without required fields. This validation ensures that the FHIR Repository will always serve sufficient data for research purposes, for example, running data mining algorithms.

Health Data Deidentification Tool

The Health Data Deidentification tool is designed to work on HL7 FHIR API so that it can be used on top of any standard FHIR

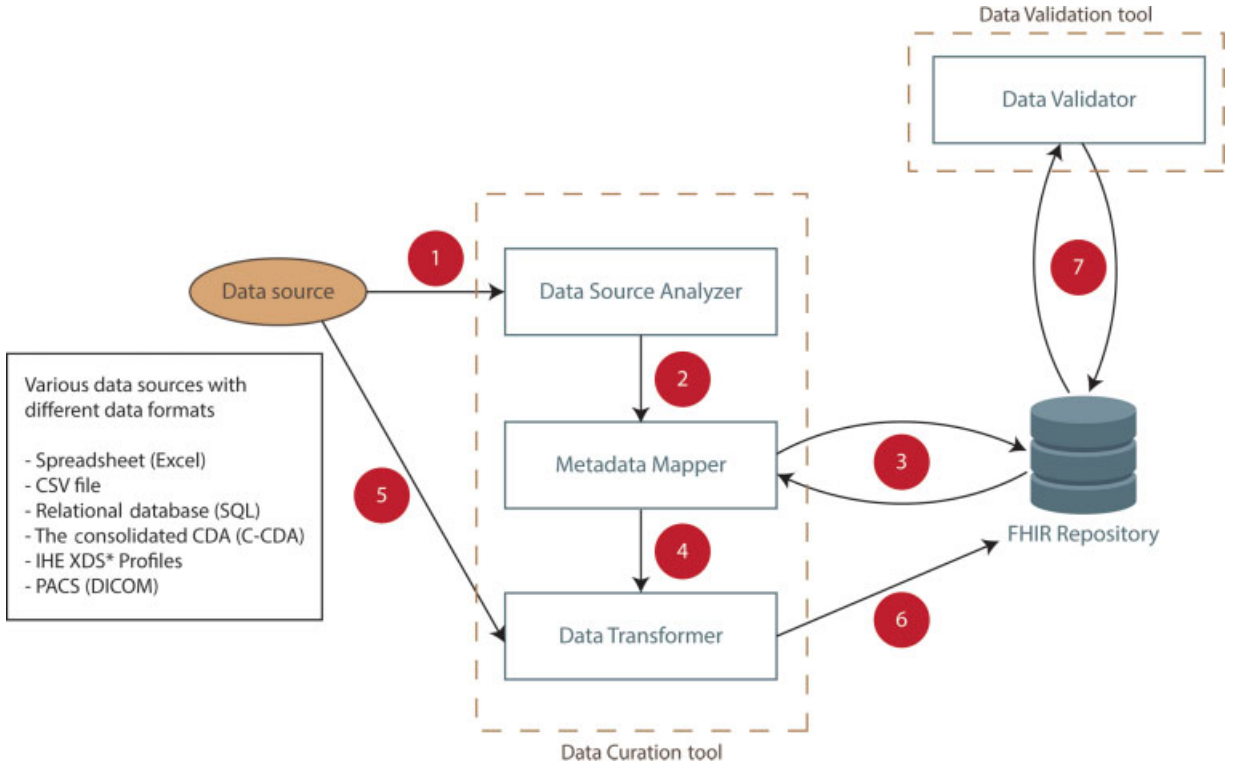

Fig. 3 The architecture of Data Curation and Validation tool. The numbered red circles show the sequence of steps explained in the text to curate and validate the health data. C-CDA, the consolidated clinical document architecture; CSV, comma-separated values; DICOM, digital imaging and communications in medicine; FHIR, fast health care interoperability resources; IHE XDS, integrating the healthcare enterprise cross-enterprise document sharing; PACS, picture archiving and communication system.

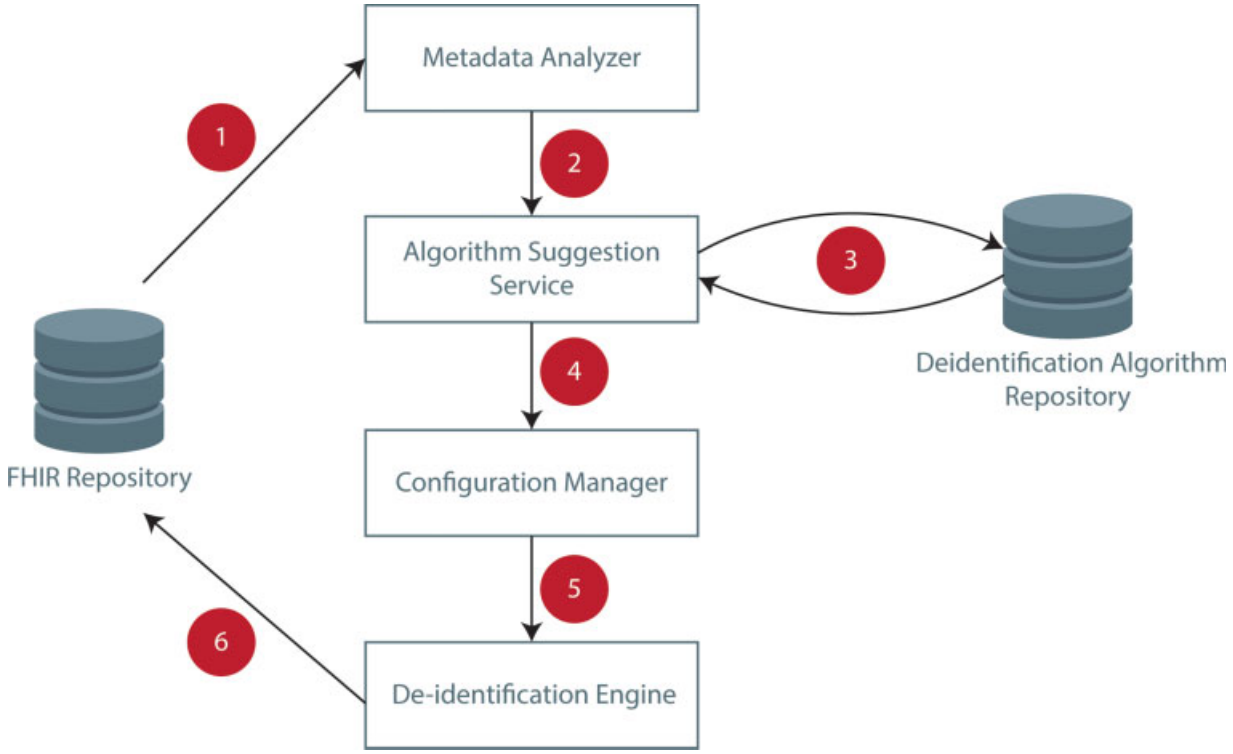

Fig. 4 The architecture of Health Data Deidentification tool. The numbered red circles show the order of health data de-identification steps that are explained in the text.

Repository as a toolset for data deidentification, anonymization and related actions. The component is expected to access FHIR resources, present metadata to the Data Manager, guide the Data Manager about the deidentification configuration to be applied and then output the processed FHIR resources. The resulting FHIR resources are deidentified/anonymized based on the configuration that the Data Manager provided.

►**Fig. 4** depicts the internal structure of Health Data Deidentification tool. It is assumed that the previous step (s) results in a FHIR Repository which contains the data as FHIR resources. Based on the presented subcomponents, a general execution scenario can be described as follows:

- Curated and validated FHIR data are analyzed by the Metadata Analyzer. In this subcomponent, available metadata are presented to the Data Manager so that the Data Manager can get detailed information about the data residing in the FHIR Repository. For example, demographic data of a patient can be described through attributes such as date of birth, gender, and nationality. Medical data observations can be described through specific attributes, for example, in case of blood pressure observation, those would be diastolic and systolic attributes.
  Algorithm suggestion service comes into play to suggest standalone algorithms (such as pseudonymization, fuzzing, and generalization) to be applied to the analyzed attributes to deidentify/anonymize. This service maintains an Algorithm Repository which includes a wide range of well-established algorithms.
- Algorithm suggestion service makes use of Deidentification Algorithm Repository to suggest algorithm. The DeIdentification Algorithm Repository is a separate component that includes different DeIdentification Algorithms. There will be common algorithms such as generalization and substitution.
- After algorithm suggestions are generated, the Configuration Manager prompts the Data Manager for the final decision. At this step, to ensure privacy preserving data publishing, the configuration manager will enable the Data Manager to run several algorithms such as k-anonymity, l-diversity, and t-closeness to assess the anonymity of the dataset, and when not satisfied, change the deidentification algorithms selected for the necessary data elements. The Data Manager will finalize configuration to select which algorithms will be used for which attributes, with which parameters.
- The Deidentification Engine receives the configurations from the Configuration Manager and connects to the FHIR Repository to convert the raw data. Data are deidentified according to those configurations.
- De-identified data are saved into the FHIR Repository again. Hence, FHIR resources which carry sensitive information are transformed into FHIR resources which do not carry sensitive information anymore.

Semantic Enrichment Tool

The Semantic Enrichment tool incorporates domain relevant medical terminologies within the FHIR CDM. It can be integrated on top of any FHIR Repository accessing the resources through the Semantic Analyzer and providing the corresponding mappings using a Terminology Service, which follows the specifications of HL7 FHIR. Finally, the FAIR data objects are stored again allowing the interpretation of

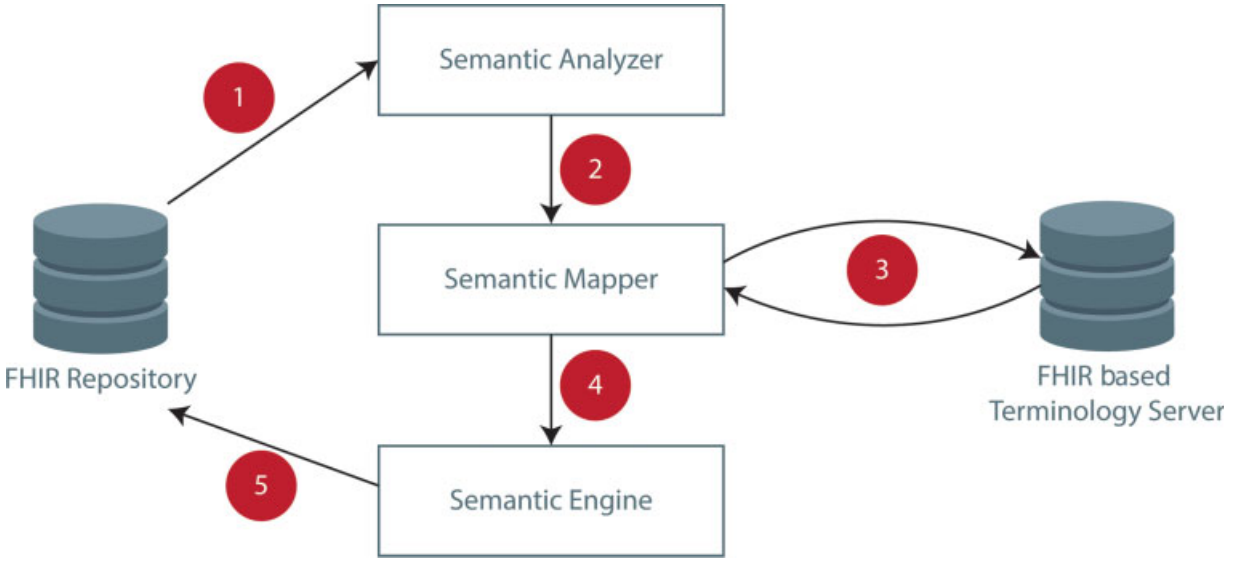

Fig. 5 The architecture of Semantic Enrichment Tool. The numbered red circles refer to the order of execution that are explained in the text. FHIR, fast health care interoperability resources.

concepts that have related meaning as if they were in the same "language."

►**Fig. 5** depicts the subcomponents of the Semantic Enrichment Tool. The workflow is as follows:

1. Curated, validated, and deidentified FHIR resources are gathered by the Semantic Analyzer. Metadata and data are retrieved and processed to identify the medical concepts and associated metadata.
2. All the resources are filtered out and those with no correspondence to the common terminology are passed to the Semantic Mapper.
3. The Semantic Mapper receives all the concepts and links the suitable translations using the FHIR-based Terminology Server which provides FHIR Concept Map,[50] resources for each of the needed translations. It also stores all the code systems (source and destination).
4. Once all the mappings are generated, the Semantic Engine fills in all the FAIR data objects with all the available translations, as the FHIR standard permits including multiple vocabularies in the same resource.
5. The data are stored back in the FHIR Repository.

Discussion

The FAIR data principles[1] aim to ensure that research outputs are shared in a way that enables and enhances reuse by humans and machines. Although FAIR emerged from a workshop in the life science community, the principles are applicable to datasets and metadata from all disciplines. FAIR echoes previous statements on open data and curation such as the Organization for Economic Cooperation and Development (OECD) principles and guidelines for Access to Research Data from Public Funding[51] and the Royal Society Science as an Open Enterprise reports.[52] The Royal Society report put forward the notion of "intelligent openness" where data are accessible, intelligible, assessable, and usable. FAIR proposed similar principles in a more arresting and memorable articulation of the concepts, and that allowed it to gain significant traction and uptake internationally. The European Commission encouraged implementation of FAIR in the call for proposals of the Horizon 2020 Work Program 2018 to 2020 under the pillar health, demographic change, and wellbeing.[53] In the same year, the NIH announced a funding opportunity for the data commons pilot phase,[54] which also supported the application of the FAIR data principles to the research outputs. A similar example can be found in Australia,[55] and this policy is already being developed in some countries of Africa, in this case with the support of the Committee on Data International Science Council (CODATA) initiative.[56] This leads to a landscape in which only those researchers able to commit to the FAIR principles will have the opportunity to develop research projects funded by public agencies. Therefore, researchers may benefit in the coming years from a common framework that supports the FAIRification process.

Considering the health-specific challenges and available research in this area, this paper proposed fine-tuned and brand-new steps within the FAIRification workflow and presented an implementation architecture. The proposed architecture utilizes an HL7 FHIR Repository for native FAIR support acting as an enabling factor for data FAIRification. FHIR specification offers modular components called resources. The characteristics of these resources can be oriented to achieve compliance with FAIR guidelines. For example, FHIR uses globally unique identifiers and can assign other identifiers. The data elements described in FAIR correspond to concepts and (meta)data objects modeled, as FHIR resources and described with rich metadata and context information. In FHIR, resources are retrievable via open APIs, that is, absolute URIs and standard Representational State Transfer (REST) protocols. All these capabilities align with FAIR principles. Steps like "Make data linkable," "License attribution," "Data versioning" and "Publishing" refer to these inherent capabilities of FHIR. Moreover, introduced steps such as "Data curation & validation" and "Data de-identification & pseudonymization" toolset

have been designed on top of a generic FHIR Repository to make it interoperable and reusable. We claim that our architectural design follows the FAIRification workflow, while implementing the revised workflow appropriate for health data.

Musen et al[57] are developing an open source workbench based on semantic web technologies to support open science within the center for expanded data annotation and retrieval (CEDAR) initiative.[58] This workbench provides a pipeline for authoring experimental metadata in biomedical sciences through the use of templates. This workbench only provides support for metadata authoring and management, and, to the best of our knowledge, it does not provide functionalities for curating, deidentifying, or versioning data. The Dutch Techcentre for Life Sciences is currently developing a set of tools to implement the FAIR data principles and apply them to the rare disease datasets collected by the RD-connect platform.[59] This toolkit is based on the FAIRification process proposed by GO FAIR, and built upon the OpenRefine software (*http://openrefine.org/*). This toolkit allows for creating, publishing, finding, and annotating FAIR datasets. However, it seems to not support the deidentification of samples, so this step would need to be performed outside the toolkit. The application of deidentification methods to the datasets may not be necessary for those samples gathered during the research process, as it is a common practice to replace personal identifiers with artificial IDs. However, when dealing with datasets derived from routine care, this step turns out mandatory so patient's privacy can be preserved.

For research purposes, data are gathered upon the signature by the data donor of an informed consent for that specific research. In a routine care environment, consent is usually given for health care purposes only, and secondary use of the personal information gathered is generally prohibited. However, GDPR acknowledges that these data can be reused without consent for reasons of public interest in the areas of public health (Art. 9). Regarding reusing this data in scientific research, GDPR permits further processing of personal data when the principle of data minimization is respected. This processing does not permit the identification of data subjects and requires that appropriate safeguards exist (such as, for instance, pseudonymization of the data) and the purpose of the processing is compatible with the original purpose for data collection (Art. 89.1, Recitals 156 and 157).[16] The FAIRification workflow proposed in this contribution supports these legal provisions as much as possible by providing the tools needed to FAIRify specific subsets of raw datasets to comply to the minimization principle and to deidentify or pseudonymize the datasets.

Conclusion

In this work, the FAIRification process proposed by GO FAIR has been examined and adapted to health data requirements by applying restrictions on existing steps and introducing new steps for requirements elicited after performing a comprehensive analysis of technical, ethical, and legal implications that reusing health data for biomedical research purposes may have. As a result of this analysis, a FAIRification workflow is proposed to be applied in the health domain taking into account-specific functionalities for data curation, data validation, data deidentification/pseudonymization, and data versioning. The technological architecture of these new components added to the FAIRification process has been designed, and the HL7 FHIR standard has been proposed for their implementation, as this standard has shown to be suitable for complying with the FAIR data principles.

Routine health care datasets or data resulting from health research can be FAIRified, shared and reused within the health research community following the FAIRification workflow and the associated architecture design proposed by this paper. This methodology leverages health data resources so that knowledge discovery can be accelerated, while reducing biases and enhancing the strength and quality of scientific evidence. Additionally, this workflow supports a practical implementation of transparency, openness, and reproducibility to support the pillars of Open Science and Responsible Research and Innovation strategies promoted by the European Commission. On top of that, the data science community will be able to take advantage of the availability of these data resources to develop advanced analytical solutions and provide data-driven innovative services that will enable a seamlessly application of new evidence into the clinical practice.

The main limitation of this contribution is that the proposed FAIRification workflow and the architecture design has not been tested in real settings yet. However, to the best of our knowledge, analyzing the health-specific requirements for various used cases, adjusting the FAIRification workflow for those specific challenges after performing thorough analyses, such as a detailed analysis of the regulatory framework in the EU, and designing a software architecture utilizing a well-established international standard is the first attempt toward methodological FAIRification of health data. Yet, it should also be noted that the workflow details and the design may be subject to change during the software development and deployment stage.

The software following the proposed architecture design is under development on Github following the open-source philosophy (https://github.com/fair4health). Future work will be focused on the practical implementation of the FAIRification workflow and on the development of used cases to demonstrate the impact that such strategy may have on health research and routine health care. More specifically, it is foreseen to develop two prototypes: the first one will support health researchers by addressing the identification of disease association patterns in the general population, while the second one will support routine healthcare through the implementation of a tool able to predict the 30-day readmission risk in patients with chronic obstructive pulmonary disease (COPD). Both used cases will be developed upon the application of this FAIRification workflow over a federated cohort accounting with more than 5 million patients based on data derived from both routine health care and health research initiatives publicly funded.

Funding

This work was performed in the scope of FAIR4Health project[31]. FAIR4Health has received funding from the

European Union's Horizon 2020 research and innovation programme under grant agreement number 824666.

Conflict of Interest

A.A.S. reports grants from EU H2020 Program, during the conduct of the study.

Acknowledgments

This research was approved by the Comité Coordinador de Ética para la Investigación Biomédica en Andalucía (Ethical Biomedical Research Steering Board of Andalusia) chaired by PhD MD Mariano Aguayo-Canela, with a file number GA82466 in Seville, September 25, 2018.

References

1 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018–160018
2 Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Inf Serv Use 2017;37:49–56
3 European Commission, Directorate-General for Research & Innovation. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. Accessed April 7, 2020
4 Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. Available at: https://eur-lex.europa.eu/eli/dir/2019/1024/oj. Accessed April 7, 2020
5 Office of Strategic Coordination—The Common Fund, National Institutes of Health. New Models of Data Stewardship. Program snapshot. Available at: https://commonfund.nih.gov/data. Accessed April 7, 2020
6 GO FAIR Initiative. Available at: https://www.go-fair.org/go-fair-initiative/. Accessed April 7, 2020
7 FAIRification process. Available at: https://www.go-fair.org/fair-principles/fairification-process/. Accessed April 7, 2020
8 Skovgaard LL, Wadmann S, Hoeyer K. A review of attitudes towards the reuse of health data among people in the European Union: the primacy of purpose and the common good. Health Policy 2019;123(06):564–571
9 Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. PLoS One 2015;10(06):e0129506–e0129506
10 Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017;26(01):38–52
11 World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013;310(20):2191–2194
12 WMA declaration of taipei on ethical considerations regarding health databases and biobanks. Available at: https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/. Accessed April 7, 2020
13 The Health Information Technology for Economic and Clinical Health (HITECH) Act Enforcement Interim Final Rule. U.S. Department of Health & Human Services, Health Information Privacy. Available at: https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.html. Accessed April 7, 2020
14 Cohen IG, Mello MM. HIPAA and protecting health information in the 21st century. JAMA 2018;320(03):231–232
15 Recommendations on de-identification of protected health information under HIPAA. U.S. Department of Health & Human Services, National Committee on Vital and Health Statistics. Available at: https://www.ncvhs.hhs.gov/wp-content/uploads/2013/12/2017-Ltr-Privacy-DeIdentification-Feb-23-Final-w-sig.pdf. Accessed April 7, 2020
16 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed April 7, 2020
17 Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc 2017;24(05):986–991
18 Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology 2016;279(02):329–343
19 Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885–h1885
20 Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. Int J Med Inform 2019;124:6–12
21 Ong T, Pradhananga R, Holve E, Kahn MG. A framework for classification of electronic health data extraction-transformation-loading challenges in data network participation. EGEMS (Wash DC) 2017;5(01):10–10
22 Hamrouni H, Brahmia Z, Bouaziz R. A systematic approach to efficiently managing the effects of retroactive updates of time-varying data in multiversion XML databases. International Journal of Intelligent Information and Database Systems 2018;11:1–26
23 Yenni GM, Christensen EM, Bledsoe EK, et al. Developing a modern data workflow for regularly updated data. PLoS Biol 2019;17(01):e3000125–e3000125
24 Wilkinson MD, Verborgh R, Bonino da Silva Santos LO, et al. Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Comput Sci 2017;3:e110
25 Collins S, Genova F, Harrower N, et al. Turning FAIR into reality. European Commission Directorate General for Research and Innovation 2018;1:1–76
26 European Commission. Ethics and data protection. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf. Accessed April 7, 2020
27 Canham S, Ohmann C, Matei M, et al. White paper 4: ethics, supporting document to D3.3 draft policy recommendations. Available at: https://eoscpilot.eu/sites/default/files/eoscpilot_d3.3_whitepaper_4_ethics.pdf. Accessed April 7, 2020
28 Ienca M, Ferretti A, Hurst S, Puhan M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. PLoS One 2018;13(10):e0204937–e0204937
29 Council of the European Union Outcome of proceedings. Available at: http://data.consilium.europa.eu/doc/document/ST-14853-2015-INIT/en/pdf. Accessed April 7, 2020
30 Floridi L, Taddeo M. What is data ethics? Philos Trans A Math Phys Eng Sci 2016;374(2083):20160360
31 FAIR4Health Project. FAIR4Health Project Website. Available at: https://www.fair4health.eu/. Accessed April 7, 2020
32 HL7 Clinical Document Architecture (CDA). Health Level Seven International (HL7). Available at: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7. Accessed April 7, 2020
33 HL7 FHIR. Available at: http://hl7.org/fhir/. Accessed April 7, 2020
34 Open industry specifications, models and software for e-health (OpenEHR). Available at: https://www.openehr.org/. Accessed April 7, 2020
35 Observational Health Data Sciences and Informatics. OMOP common data model. Available at: https://www.ohdsi.org/data-standardization/the-common-data-model/. Accessed April 7, 2020

36 U.S. Department of Health & Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. Accessed April 7, 2020

37 Burrows JH. Secure hash standard. In: Federal Information Processing Standards Publication. National Institute of Standards and Technology; 1994

38 Lakshmanan T, Madheswaran M. A novel secure hash algorithm for public key digital signature schemes. Int Arab J Inf Technol 2012;9:262–267

39 Dalenius T. Finding a needle in a haystack or identifying anonymous census records. J Off Stat 1986;2:329–336

40 World Health Organization. Classification of Diseases (ICD)-11. Available at: https://www.who.int/classifications/icd/en/. Accessed April 7, 2020

41 SNOMED International. Available at: http://www.snomed.org/. Accessed April 7, 2020

42 LONIC. The international standard for identifying health measurements, observations, and documents. Available at: https://loinc.org/. Accessed April 7, 2020

43 Rauber A, Asmi A, van Uytvanck D, Proell S. Data citation of evolving data: Recommendations of the Working Group on Data Citation (WGDC). Result of the RDA Data Citation WG 2015;20:1–2

44 ISO 14721:2003 Space data and information transfer systems—open archival information system—Reference model. Available at: https://www.iso.org/standard/24683.html. Accessed April 7, 2020

45 Deserno TM, Welter P, Horsch A. Towards a repository for standardized medical image and signal case data annotated with ground truth. J Digit Imaging 2012;25(02):213–226

46 Canham S, Ohmann C. A metadata schema for data objects in clinical research. Trials 2016;17(01):557–557

47 Cross-Enterprise Document Sharing. Available at: https://wiki.ihe.net/index.php/Cross-Enterprise_Document_Sharing. Accessed April 7, 2020

48 Digital Imaging and Communications in Medicine (DICOM). Available at: https://www.dicomstandard.org/. Accessed April 7, 2020

49 C-CDA (HL7 CDA R2 Implementation Guide: Consolidated CDA Templates for Clinical Notes—US Realm). Available at: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=492. Accessed April 7, 2020

50 FHIR HL7. Resource ConceptMap—Content. Available at: https://www.hl7.org/fhir/conceptmap.html. Accessed April 7, 2020

51 OECD Principles and Guidelines for Access to Research Data from Public Funding. Available at: http://www.oecd.org/sti/inno/38500813.pdf. Accessed April 7, 2020

52 The Royal Society. Science as an open enterprise: open data for open science. Available at: https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf. Accessed April 7, 2020

53 European Commission. Horizon 2020, Work Programme 2018–2020, Health, demographic change and wellbeing. Available at: https://ec.europa.eu/programmes/horizon2020/sites/horizon2020/files/health_h2020_draft_sc1_wp_18-20_0.pdf. Accessed April 7, 2020

54 Notice Announcing Funding Opportunity Issued for the NIH Data Commons Pilot Phase. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-RM-17-031.html. Accessed April 7, 2020

55 2016 National Research Infrastructure Roadmap. Available at: https://docs.education.gov.au/system/files/doc/other/ed16-0269_national_research_infrastructure_roadmap_report_internals_acc.pdf. Accessed April 7, 2020

56 University of Nebraska–Lincoln. The african open science platform: the future of science and the science of the future. Available at: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1092&context=scholcom. Accessed April 7, 2020

57 Musen MA, Sansone S-A, Cheung K-H, et al. CEDAR: Semantic Web Technology to Support Open Science. In WWW'18Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France. 2018;2:427428

58 Musen MA, Bean CA, Cheung KH, et al; CEDAR team. The center for expanded data annotation and retrieval. J Am Med Inform Assoc 2015;22(06):1148–1152

59 Thompson M, Bonino L, Wilkinson MD, et al. Overview of a suite of middle-ware services for implementing FAIR data principles. CEUR Workshop Proc 2017