Related to other papers in this special issue	4 (p40); 12 (p122); 23 (p230); 19 (p192); 5 (p47); 20 (p199); 9 (p87); 7 (p66)
Addressing FAIR principles	F, A, I, R

Distributed Analytics on Sensitive Medical Data: The Personal Health Train

Oya Beyan[1,2†], Ananya Choudhury[3], Johan van Soest[3,4], Oliver Kohlbacher[5,6,7,8], Lukas Zimmermann[7], Holger Stenzhorn[7], Md. Rezaul Karim[1,2], Michel Dumontier[4], Stefan Decker[1,2], Luiz Olavo Bonino da Silva Santos[9] & Andre Dekker[3]

[1]Fraunhofer Institute for Applied Information Technology (FIT), 53754 Sankt Augustin, Germany

[2]RWTH Aachen University, 52056 Aachen, Germany

[3]Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center, 6200 MD Maastricht, The Netherlands

[4]Institute of Data Science, Maastricht University, Universiteitssingel 60, Maastricht 6229 ER, The Netherlands

[5]Department of Computer Science, University of Tübingen, Tubingen, Baden-Württemberg 72076, Germany

[6]Quantitative Biology Center, University of Tübingen, Tubingen, Baden-Württemberg 72076, Germany

[7]Institute for Translational Bioinformatics, University of Tübingen, Tubingen, Baden-Württemberg 72076, Germany

[8]Center for Bioinformatics, University of Tübingen, Germany

[9]GO FAIR International Support & Coordination Office (GFISCO), Leiden, The Netherlands

Keywords: Distributed analytics; Data reuse; FAIR; Health data; Ethics and privacy

Citation: O. Beyan, A. Choudhury, J van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, Md. R. Karim, M. Dumontier, S. Decker, L.O. Bonino da Silva Santos & A. Dekker. Distributed analytics on sensitive medical data: The Personal Health Train. Data Intelligence 2(2020), 96–107. doi: 10.1162/dint_a_00032

ABSTRACT

In recent years, as newer technologies have evolved around the healthcare ecosystem, more and more data have been generated. Advanced analytics could power the data collected from numerous sources, both

† Corresponding author: Oya Beyan (E-mail: beyan@fit.fraunhofer.de, ORCID: 0000-0001-7611-3501).

from healthcare institutions, or generated by individuals themselves via apps and devices, and lead to innovations in treatment and diagnosis of diseases; improve the care given to the patient; and empower citizens to participate in the decision-making process regarding their own health and well-being. However, the sensitive nature of the health data prohibits healthcare organizations from sharing the data. The Personal Health Train (PHT) is a novel approach, aiming to establish a distributed data analytics infrastructure enabling the (re)use of distributed healthcare data, while data owners stay in control of their own data. The main principle of the PHT is that data remain in their original location, and analytical tasks visit data sources and execute the tasks. The PHT provides a distributed, flexible approach to use data in a network of participants, incorporating the FAIR principles. It facilitates the responsible use of sensitive and/or personal data by adopting international principles and regulations. This paper presents the concepts and main components of the PHT and demonstrates how it complies with FAIR principles.

1. INTRODUCTION: MOVING FROM CENTRALIZED DATA SHARING TO EMPOWERING DATA OWNERS TO GAIN CONTROL OVER DATA REUSE

Data-driven technologies are changing business, our daily lives, and the way we conduct research more than ever. In recent years, more and more data have been generated in the healthcare ecosystem. The data contain potential knowledge to transform health care delivery and life sciences. Advanced analytics could potentially power the data collected from numerous sources to improve prevention, diagnosis and treatment of diseases, as well as supporting individuals and societies to maintain their health and well-being.

The era of exponential growth of data has also witnessed the increase of risk involved in sharing them. Countries are quickly adopting policies and formulating laws that regulate the collection, use, and sharing of personal data. The data protection law in the USA, the HIPAA Act, limits sharing sensitive data. In the European Union, the General Data Protection Regulation sets a well-formulated directive for securing confidentiality and privacy of citizens so that the data are not available publicly without explicit, well informed specific consent, and cannot be used to identify a subject without additional information stored separately [1]. PIPEDA in Canada, the Data Protection Act (PDA) in the UK, the Russian Federal Law on Personal Data, the IT Act in India and the China Data Protection Regulations (CDPR), all reflect the increasing global awareness regarding the importance of data privacy and confidentiality [2, 3, 4, 5, 6]. Patients and the general public are becoming more and more aware about the use of their personal data and are becoming more reluctant about sharing data. The current norm is that disclosure of health data without proper consent is a breach of privacy, which harms the fundamental right of freedom from intrusion or interference by others. Organizations that safeguard trusted information have thus a duty to ensure confidentiality [7]. Alternatively, anonymization and data masking are common solutions applied to protect privacy, although these methods cannot totally mitigate the risk of re-identification [8]. Big data analytics applications increase the risk of (re)identification, since linking various data sources increases the amount and quality of information [3, 9]. These high dimensional data sets can be used to infer sensitive information at the individual or at the subpopulation level.

Due to ethical concerns, a huge amount of usable health data is currently trapped inside the organizational boundaries of hospitals, clinics and within patients' devices. Many healthcare institutions implement centralized repositories by pooling data from multiple systems into data warehouses or data lakes [10]. Sharing these data out of the organization's boundaries is not a viable solution since the anonymization of data may not be possible for certain data types such as genomic data and also since linking data sets increases the re-identification risk. Alternatively, research communities build domain-specific data infrastructures [11] e.g., bioinformatics, cohort studies, clinical research or biobanks. The problem of accessing data outside of the network remains, and since data are collected for a specific use and duplicated outside of the first data source, it limits the record linkage and integration of multimodal data.

As a technical solution to centralized data sharing, i2b2 or DataShield provide software and tools to support querying and analysis of sensitive data in a distributed fashion by proposing their own technology stack and tools [12, 13, 14]. Nonetheless, since health data are generated and stored in a highly diverse system by heterogeneous stakeholders, it is very unlikely that these infrastructures will converge on a single solution.

Another aspect is social and cultural. The sensitive nature of health data makes individuals and institutions hesitant to share their data. From the public perspective, people are more willing to accept and participate in data sharing if they are informed about existing safeguards and governance mechanisms. They are willing to contribute to science for better care and wellbeing; however, they want to decide who can use their data, for what purposes and make sure data users are accountable for their actions. A survey among 603 secondary data users shows that 56% of the researchers who are willing to share their data demand a context with access control, and want to have a say or at least knowledge regarding the use of the data [15]. The current data sharing practice does not allow the owners to decide who can access the data and for which purpose. Although data sharing and licensing agreements set terms and conditions such as the limitation to a certain number of research purposes, the conditions of data transfer, not allowing attempts to establish individual identities or a maximally allowed time before data have to be destroyed, once data is out of the institutional boundaries, there is no mechanism for enforcement of these policies.

The Personal Health Train (PHT) proposes an alternative approach which encompasses both technological and social aspects of sensitive data reuse. When data sharing is not achievable, using distributed analytics on distributed data becomes a viable solution. The PHT does not require the transfer of data from the holding entity. Rather than moving the data to the requester, it moves the analytics tasks to the data repositories and executes the tasks in a secure environment. In this approach, the owner of the data can remain in control and decide which part of the data will be analyzed for which specific purposes and by whom. This new approach requires discovering, understanding, exchanging and executing analytics tasks with minimum human intervention.

FAIR principles becomes relevant not only for data but also for analytics tasks. In the fragmented landscape of data, interoperability and accessibility can be ensured by applying FAIR principles to the analytics tasks

and system components that interact with these tasks. In this paper, we will demonstrate the application of FAIR principles to the Personal Health Train approach.

2. AN OPEN ECOSYSTEM WHERE DATA MEETS ANALYTICS: MACHINE READABILITY AT THE CORE

The Personal Health Train provides an infrastructure to support distributed and federated solutions that utilize the data at the original location. Typically health data are produced by diverse sources, including care institutions, biomedical researchers, imaging facilities, clinical and population studies, genomic sequencing centers and by citizens themselves. It creates an open ecosystem by making self-contained, machine-readable analytics task exchangeable and executable in diverse systems. The PHT does not prescribe any specific standard or technology for data, and instead, it only requires publishing individual choices as metadata. The PHT focuses on making data, tasks, processes and algorithms findable, accessible, interoperable and reusable (FAIR). As a result, it enables data providers and data users to match FAIR data to FAIR analytics and empowers them to make informed decisions about participating in specific applications.

The PHT provides an alternative solution to reuse the data in institutional data silos or citizens' personal data stores. It targets maximal interoperability between diverse systems, by focusing on machine-readable and interpretable data, metadata, workflows and services. The core design principle is to give data owners authority to decide and monitor the use of their data. Eventually, this will lead to the creation of the Internet of FAIR data and services that operates on personal health data that can never be completely open.

An example application is training of patient surviving prediction model. The particular case requires to assess and analyze a large amount of real-world, high-dimensional, multimodal personal data. In the health domain, this corresponds to information such as longitudinal medical records, diagnostics tests such as imaging, genomic profiles, and patient-generated health data and outcomes via apps and wearable devices. To discover hidden patterns, the full data set should be made available to the machine learning task, but the privacy-driven requirement of data minimization limits the personal data to those elements deemed directly relevant and necessary to accomplish a specified purpose. The PHT approach could unleash the potential of big data analytics for personal data without compromising privacy. The machine learning model can be sent across various health-care providers through the PHT infrastructure without data ever leaving the organizational boundaries [16, 17].

The PHT defines the following three core components:

Station: Provides curated, confidential data and acts as FAIR data points. Stations expose data in a discoverable format, define an interface to execute queries, provide computational resources and execute analytic tasks in a secure environment. Stations are registered and the schemas and the metadata of the data provided by a Station are published through *Station Registries*.

Train: Data Consumers intend to access privacy-sensitive data from multiple curators and to execute a data analytics algorithm to derive insights from the data. They formulate the queries and specify the analytics

algorithm. The set of all artifacts required to execute the distributed algorithm and return the results is called a "Train". A Train is identified by a Digital Persistent Identifier (PID) and contains a self-sufficient message with all the information required to transfer code and result between the relevant parties. Trains may be simple or complex with different kinds of wagons that are also digital uniquely identifiable objects. Each wagon may have its own resources with many different types. A Train carries different components; namely, metadata that stores the Train's unique digital persistent identifier, study description, the query used in data extraction, analytics for data utilization and aggregation for result integration. Once specified, the consumer uploads the Train to the *Train Repository* and sends the reference of the Train to the handling Station. Trains are registered in a *Train Registry* to make them identifiable. The consumer has no direct access to the data sources and humans are entirely decoupled from the computation phase until the algorithm has finished.

Handler (Track): Acts as a gateway between the consumer and the curators. It orchestrates communication by receiving self-sufficient Trains from the consumer and forward them to selected Stations. It may act as a broker and may aggregate results from multiple curators. It manages Train and Station states and logs the transaction information for future auditing. Essentially, the Track is a centralized point of trust. The Train dispatcher module of the Track transfers the PHT Train either as payload or as a reference. Container execution modules at the Station (platforms in the PHT metaphor) consume the PHT Train and execute the provided algorithm. The results from different Stations are evaluated and aggregated by the Track and sent back to the consumer.

The PHT proposes a technology agnostic implementation by definition of a commonly agreed Train metadata. By design, it enables shipment of any analytic task written in any programming language. Figure 1 sketches a high-level representation of the various components of the PHT architecture.

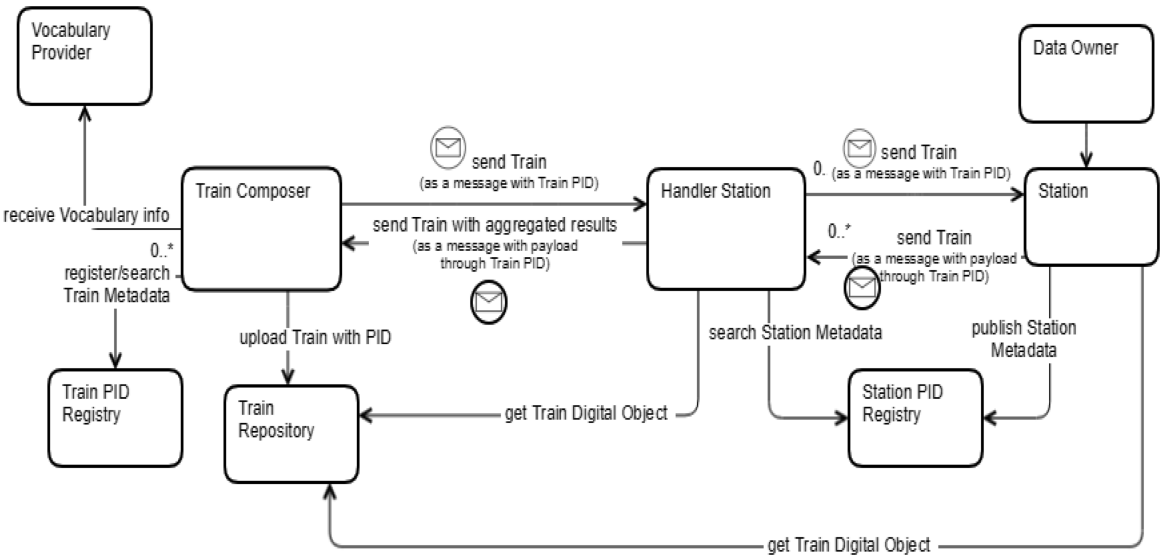

Figure 1. Main components of the PHT architecture.

3. FOLLOWING FAIR PRINCIPLES FOR DISTRIBUTED ANALYTICS

FAIR refers to a set of guiding principles that aims to enhance the ability of machines and individuals to automatically find and use data [18, 19]. Although it is originally designed for data management and stewardship with a focus on making data self-explainable and discoverable, it can be applied to any digital object with a goal to create an integrated and harmonized domain to support reusability [20]. The PHT approach promotes improving the reuse of data by sharing analytics, which can interact with the data and complete its task without giving access to the end user. Within the PHT, the FAIR principles are applied to both the Train and Station concepts, keeping in mind that the goal is enhancing the reusability of distributed data with distributed analytics.

Clearly, making data self-explainable and discoverable goes a long way to ensure reusability. However, this may not always be possible, specifically when data are sensitive and have not been collected for research purposes. In the case of data collected during routine healthcare, for example, it is likely that data are stored in heterogeneous systems and follow the data standards imposed by the requirements of daily transactions, such as HL7 or DICOM, which might not support the desired level of metadata and persistent identification schemes. Therefore, the PHT needs to interact with data repositories, which may or may not follow FAIR principles, despite the fact that having FAIR data is highly desirable. Participating data repositories independently decide at which degree they will support FAIR data. They act as FAIR data points [21] by implementing custom interfaces supporting the computational task that reuses data.

The PHT sets the machine readability at the core, aiming for maximal interoperability between diverse systems. Therefore, it is well aligned with FAIR principles. The components of the PHT infrastructure support FAIR principles at varying degrees (Figure 2).

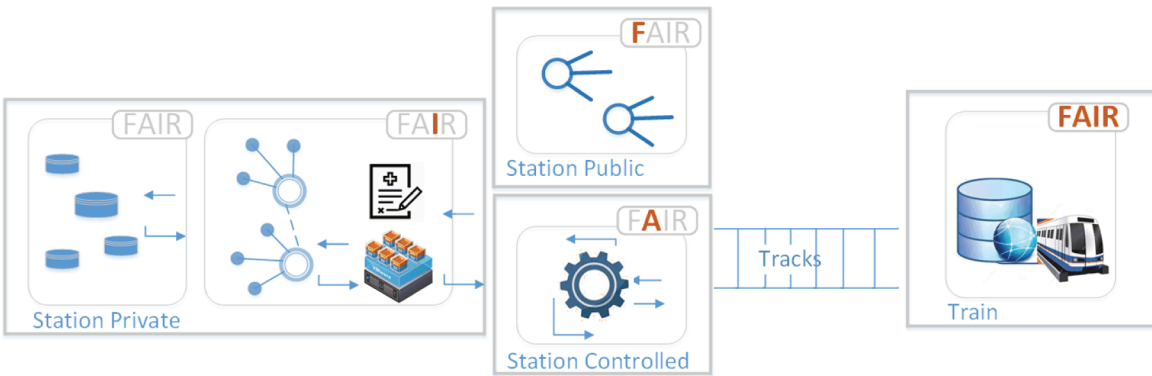

Figure 2. Applicability of FAIR principles to the components of the PHT.

Station Private: Access to health data has restrictions which derive from the original consent obtained by the patient or from data protection policies of involved institutions [22]. These data should be kept in a secure part of the Station which is not accessible by external data consumers. FAIR is not a requirement

for the private part of data repositories data and metadata reside in, which may follow preferred institutional standards. However, the access of the analytics tasks should be supported by having a queriable consent, a mechanism to link data sets, and a virtual layer to support integrated queries over diverse data sets. Therefore, it requires applying a formal and shared knowledge representation. In conclusion, the private part of the Station should support ***Interoperability***.

Station Controlled: This part of the Station provides a secure environment for executing analytics tasks. It supports ***Accessibility*** by following standardized communication protocols to discover and receive Trains. Analytics tasks can be delivered with open, free and universally implementable protocols mandating authentication and authorization procedure. Access control to data resides in the sovereignty of the Station, but results are communicated with open protocols. Ideally, ontology-based access control can be applied [23].

Station Public: Each Station is uniquely identified with a persistent identifier and registered in a registry with its metadata. It improves findability by publishing the metadata about the data repositories, as well as the computational environment.

Trains: Data analytics tasks support all four dimensions of FAIR metrics. They are ***Findable***, as Trains are uniquely and persistently identified resolvable digital objects that are registered in a Train Registry, searchable by their metadata. Train objects are persistently stored in repositories that contain all source and environment information required to

What FAIRness means for a PHT Station:

- (F) As a data owner, I want to provide enough metadata to be discoverable and published by Station registry;
- (I+A) As a Station administrator, I want to judge if a specific Train can use my data (e.g. compatible data standards), or if I have the required computational resources (e.g., metadata descriptions of Trains) before I provide a permission;
- (I+A) As a Station administrator/dispatcher, I want to set a mechanism to prevent a high demand for computational resources (e.g., prevent a crash in the Station);
- (I+A) As a Station administrator, I want to interact with Trains through defined interfaces for providing data input, and executing the tasks.

What FAIRness means for a PHT Train:

- (F) As a data consumer, I want to find already implemented Trains for a specific task (e.g. calculate hospital readmission rates for a specific case) (Train metadata);
- (F+R) As a data consumer/owner, I want to find exactly the same Train without any change after two years to replicate the computation (persistency policy for Trains);
- (A+I) As a data owner, I want to guarantee that the Train deposited and persistently identified in a repository, is the same Train that I receive (e.g. methods such as checksum);
- (A) As a data consumer, I want to guarantee that the Train that I am sending over public network is securely transferred (is there a mechanism e.g. some public/private keys);
- (A) As a data consumer/owner, I want to apply authorization and authentication policies to Train repositories for identity management.

execute them. They are ***Accessible*** with open, free, and universally implementable protocols allowing authentication and authorization. They are ***Interoperable*** since every Train described by metadata uses a formal, accessible, shared and broadly applicable language, e.g., XML, for knowledge representation. The metadata defines both the content and provenance of the analytics task such as what is the intended use, who developed it, what are the consent requirements, and also the requirements of specific tasks such as dependencies, prescribed data standards and computational resources. They are self-contained which enable virtualization to support interoperability during execution. Trains are ***Reusable***, they are designed to be reused in multiple locations. License and certification can be assigned to Trains. They keep detailed provenance metadata, including execution history.

Train repositories are the building stones to achieve the FAIRness of analytics tasks. They should adopt and follow the recommendations set for the data repositories, namely persistent identification, application programming interface (API), Train curation and moderation workflows, accessibility, license for reuse, and sustainability [24]. The first recommendation to follow is assigning persistent and global identifiers to each Train. Various identifier schemas such as URIs or DOIs can be employed. Trains which are deposited to private registries should be described with rich descriptive and operational metadata and can be registered to public repositories such as DataCite[①]. Trains should receive a PID ideally at the earliest workflow state and in order to support later operations the PID should be embedded to the object [25]. Identification of Trains with PIDs and having associated machine-readable metadata can facilitate distribution of Trains in a Digital Object Architecture [26]. The second recommendation for FAIR Train repositories is to offer a set of well-documented APIs to ensure programmatic access to Trains and Train metadata. The next recommendation is providing a platform to support data scientists to define and moderate their Trains composed of analytics tasks and metadata. Similar to data curation experts, data scientists will require tools where they can check, verify and approve the content. The accessibility requirement of the Train repository should be ensured by open and implementable protocols such as HTTP(S) and FTP. Moreover, licenses for reuse should be clearly defined for Trains. Currently, there are various options for licensing data and database

What FAIRness means for PHT Tracks:

- As a dispatcher, I want to have enough metadata to understand the content and requirements of a Train, so that I can route them to the relevant Stations;
- As a dispatcher, I want to have enough metadata about Stations, so that I can communicate with them and direct the relevant Trains to matching Stations;
- As a dispatcher/auditor, I want to check who is sending this specific Train, for which purposes and to which data Stations;
- As a dispatcher, I want to have enough information about the status of the computation, so that I can communicate results when an execution step is finished;
- As an auditor, I want to have enough provenance metadata so that I can trace and replicate execution flows when needed.

① https://datacite.org/.

rights [27]. Further investigation should be carried out to associate licenses to Trains reflecting the intellectual property and copyrights of analytics task. The last requirement is sustainability: Train repositories should have a long term preservation strategy.

The **PHT Track or Handler** monitors the request/response cycle between Trains and Stations and executes the aggregation tasks whenever required. All communication is logged by the Handler. As a result, it improves the transparency and accountability of the involved partners. Table 1 summarizes the supported FAIR principles by the PHT.

Table 1. FAIR principles supported by the PHT components.

PHT Concepts	Functionalities	FAIR Principles
Trains	They are data analytics tasks that are uniquely identified, richly described with metadata, registered and deposited to repositories. They are machine readable and executable digital objects.	Findable Accessible Interoperable Reusable
Station Private	Contains private data repositories and a data integration layer. Links data, stores access rights, exposes data with a standard representation by using terminologies and vocabularies.	Interoperable
Station Controlled	Executes Trains in a controlled environment. It has defined protocols to communicate with Trains and must also have authentication and authorization procedures. Log data are available after the execution of the task is complete.	Accessible
Station Public	Stations are registered in a repository with metadata. They publish both metadata about the contained data repositories and computational capabilities.	Findable
Track	Provides communication protocols and keeps track of all the communication. Supports traceability and reproducibility of the executed analytics.	beyond FAIR

4. CONCLUSION AND OUTLOOK

The PHT is a novel approach establishing a FAIR distributed data analytics infrastructure enabling the (re)use of distributed healthcare data, while data owners stay in control of their own data. In summary the PHT:

(i) empowers citizens and organizations to control the use of the data that reside in their own data repositories for the benefit of the individual and society,

(ii) improves the usability of health data by lowering the barriers for data protection, by ensuring that the privacy and confidentiality of the data subject will be preserved,

(iii) ensures data sovereignty beyond data security and privacy by supporting the responsible use and builds trust between data consumers and data owners by making analytics processes repeatable, transparent and auditable,

(iv) applies FAIR principles to the protocols of how data analytics interacts with FAIR data points by making data analytics tasks itself FAIR and placing machine readability at its core.

The PHT provides a distributed, flexible approach to use data in a network of participants, incorporating the FAIR principles. The PHT facilitates the responsible use of sensitive and/or personal data by adopting international principles and regulations. It supports accountability by providing provenance of analytics execution and audit mechanisms.

The PHT has been already implemented in various use cases. The Maastro clinic has implemented a Patient Cohort Counter (PCC) "Train" as a demonstration using multiple data representations. The PCC calculates the number of matching patients and cohort statistics for a specific disease at a PHT data Station. The PCC can work with different data sources with different data representations (e.g., FHIR, RDF, OMOP-OHDSI, CDISC-ODM) and is agnostic to the underlying data. The current implementation works with two data sources, one with RDF based on the Radiation Oncology Ontology, and one Station using FHIR[2] Other applications are the Varian Learning Portal by Varian Medical Systems and the open source software ppDLI by IKNL which are both example implementations of distributed learning PHT infrastructures in healthcare[3]. One use case demonstration is the development of a distributed Bayesian network model to predict dyspnea after radiotherapy for lung cancer patients which has been developed and used in the Varian Learning portal using data from five different hospitals. The ppDLI implementation currently provides a ready to use implementation of distributed Cox Proportion Hazards algorithm [16]. SMITH and DIFUTURE projects funded by the German Medical Informatics Initiative have developed cross consortia implementations and tested phenotyping use cases [28]. The PHT approach can be applied to various other domains, which want to process data but cannot share them due to the sensitive nature of data, such as the agricultural sector and the courts.

AUTHOR CONTRIBUTIONS

O. Beyan (beyan@fit.fraunhofer.de) conceived and designed the concept and wrote the paper. A. Choudhury (ananya.choudhury@maastro.nl) wrote the manuscript, and is developing the infrastructure. J. van Soest (johan.vansoest@maastro.nl) reviewed the manuscript and is working on PHT infrastructure development and implementations. O. Kohlbacher (oliver.kohlbacher@uni-tuebingen.de) reviewed the paper and and conceived core components of the architecture. L. Zimmermann (lukas.zimmermann@uni-tuebingen.de) reviewed the paper and works on components of the PHT architecture. H. Stenzhorn (holger.stenzhorn@uni-tuebingen.de) reviewed the paper and participates in the PHT development. R. Karim (rezaul.karim@fit.fraunhofer.de) reviewed the manuscript and is working on PHT infrastructure development and implementations. M. Dumontier (michel.dumontier@maastrichtuniversity.nl) conceived and reviewed the paper. S. Decker (stefan.decker@fit.fraunhofer.de) and A. Dekker (andre.dekker@maastro.nl) conceived

[2] https://github.com/jvsoest/PHT_on_FHIR_demo.
[3] https://distributedlearning.ai/blog/.

and reviewed the paper. L.O. Bonino da Silva Santos (luiz.bonino@go-fair.org) reviewed the paper and works on the design of the Personal Health Train architecture.

REFERENCES

[1] General Data Protection Regulation (GDPR). Available at: https://gdpr-info.eu/.

[2] Office of the Privacy Commissioner of Canada. The Personal Information Protection and Electronic Documents Act (PIPEDA). Available at: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/.

[3] The Data Protection Act. Available at: https://www.gov.uk/data-protection.

[4] Federal Law of 27 July 2006 N 152-FZ on Personal Data. Available at: https://pd.rkn.gov.ru/authority/p146/p164/.

[5] Ministry of Electronics and Information Technology, Government of India. Information Technology Act. Available at: https://meity.gov.in/content/information-technology-act.

[6] China Data Protection Regulations (CDPR). Available: https://www.chinalawblog.com/2018/05/china-data-protection-regulations-cdpr.html.

[7] Privacy and confidentiality: The interagency advisory panel on research ethics (PRE). Available at: http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/chapter5-chapitre5/.

[8] K. El Emam, S. Rodgers & B. Malin. Anonymising and sharing individual patient data. BMJ 350(2015), h1139. doi: 10.1136/bmj.h1139.

[9] V. Torra & G. Navarro-Arribas. Big data privacy and anonymization. In: A. Lehmann et al. (eds.) Privacy and Identity Management. Facing up to Next Steps. Privacy and Identity 2016. Cham, Switzland: Springer. doi: 10.1007/978-3-319-55783-0_2.

[10] Secondary use of clinical data: The Vanderbilt approach. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24534443.

[11] A distributed infrastructure for life-science information. Available at: https://elixir-europe.org/.

[12] i2b2 Research Data Warehouse. Available at: https://i2b2.cchmc.org/.

[13] DataSHIELD - Newcastle University. Available at: http://www.datashield.ac.uk/about/howdoesdatashieldwork/examplesofdatashieldinfrastructure/.

[14] DataSHIELD – New directions and dimensions. Available at: https://datascience.codata.org/articles/10.5334/dsj-2017-021/.

[15] What drives academic data sharing? Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118053.

[16] A. Jochems, T.M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, ... & A. Dekker. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Clinical and Translational Radiation Oncology 121(3)(2016), 459–467. doi: 10.1016/j.radonc.2016.10.002.

[17] T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, ... & P. Lambin. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and Translational Radiation Oncology 4(2017), 24–31. doi: 10.1016/j.ctro.2016.12.004

[18] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[19] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O. Bonino da Silva Santos & M.D. Wilkinson. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37(2017), 49–56. doi: 10.3233/ISU-170824.

[20] P. Wittenburg, F. de Jong, D. van Uytvanck, M. Cocco, K. Jeffery, M. Lautenschlager, ... & P. Holub. State of FAIRness in ESFRI projects. Data Intelligence 2(2020), 230–237. doi: 10.1162/dint_a_00045.

[21] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos & L.O. Bonino da Silva Santos. Making FAIR easy with FAIR tools: From creolization to convergence. Data Intelligence 2(2020), 87–95. doi: 10.1162/dint_a_00031.

[22] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The "A" of FAIR – as open as possible, as closed as necessary. Data Intelligence 2(2020), 47–55. doi: 10.1162/dint_a_00027.

[23] C. Brewster, B. Nouwt, S. Raaijmakers & J. Verhoosel. Ontology-based access control for FAIR data. Data Intelligence 2(2020), 66–77. doi: 10.1162/dint_a_00029.

[24] M. Hahnel & D. Valen. How to (easily) extend the FAIRness of existing repositories. Data Intelligence 2(2020), 192–198. doi: 10.1162/dint_a_00041.

[25] T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. Data Intelligence 2(2020), 40–46. doi: 10.1162/dint_a_00026.

[26] L. Lannom, D. Koureas & A.R. Hardisty. FAIR data and services in biodiversity science and geoscience. Data Intelligence 2(2020), 122–130. doi: 10.1162/dint_a_00034.

[27] I. Labastida & T. Margoni. Licensing FAIR data for reuse. Data Intelligence 2(2020), 199–207. doi: 10.1162/dint_a_00042.

[28] Md. R. Karim, B.P. Nguyen, L. Zimmermann, T. Kirsten, M. Löbe, F. Meineke, ... & O. Beyan. A distributed analytics platform to execute FHIR-based phenotyping algorithms. Available at: http://ceur-ws.org/Vol-2275/paper8.pdf.