Distributed Analytics on Sensitive Medical Data: The Personal Health Train (Postprint)

Authors: Beyan, Oya, Choudhury, Ananya, van Soest, Johan, Kohlbacher, Oliver, Zimmermann, Lukas, Stenzhorn, Holger, Karim, Md Rezaul, Dumontier, Michel, Decker, Stefan, Santos, Luiz Olavo Bonino da Silva, Dekker, Andre, Beyan, Oya

Date: 2022-11-16T00:00:00+00:00

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 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.

Full Text

Preamble

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, newer technologies evolving around the healthcare ecosystem have generated increasingly large volumes of data. Advanced analytics could power data collected from numerous sources—both from healthcare institutions and generated by individuals via apps and devices—to drive innovations in disease treatment and diagnosis, improve patient care, and empower citizens to participate in decision-making about their own health and well-being. However, the sensitive nature of health data prohibits healthcare organizations from sharing it. The Personal Health Train (PHT) is a novel approach that establishes a distributed data analytics infrastructure enabling the (re)use of distributed healthcare data while data owners retain control of their own data. The main principle of the PHT is that data remain in their original location while analytical tasks visit data sources and execute on-site. The PHT provides a distributed,

flexible approach to using 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 its compliance with FAIR principles.

1. Introduction: Moving from Centralized Data Sharing to Empowering Data Owners to Gain Control Over Data Reuse

Data-driven technologies are transforming business, daily life, and research conduct more than ever before. In recent years, the healthcare ecosystem has generated increasingly large volumes of data containing knowledge with the potential to transform health care delivery and life sciences. Advanced analytics could power data collected from numerous sources to improve disease prevention, diagnosis, and treatment, while supporting individuals and societies in maintaining health and well-being.

The era of exponential data growth has also witnessed increased risks involved in sharing such data. Countries are rapidly adopting policies and formulating laws that regulate personal data collection, use, and sharing. In the USA, the HIPAA Act limits sharing of sensitive data. In the European Union, the General Data Protection Regulation establishes a well-formulated directive for securing citizen confidentiality and privacy, ensuring data are not publicly available without explicit, well-informed specific consent and cannot be used to identify subjects without separately stored additional information [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 growing global awareness regarding data privacy and confidentiality [2, 3, 4, 5, 6].

Patients and the public are becoming increasingly aware of how their personal data is used and more reluctant to share it. The current norm holds that disclosure of health data without proper consent constitutes a privacy breach that harms the fundamental right to freedom from intrusion or interference. Organizations safeguarding trusted information thus have a duty to ensure confidentiality [7]. While anonymization and data masking are common privacy protection solutions, these methods cannot fully mitigate re-identification risk [8]. Big data analytics applications increase (re)identification risk because linking various data sources enhances the amount and quality of information [3, 9]. These high-dimensional datasets can be used to infer sensitive information at individual or subpopulation levels.

Due to ethical concerns, vast amounts of usable health data remain trapped within organizational boundaries of hospitals and clinics or on patients' per-

sonal devices. Many healthcare institutions implement centralized repositories by pooling data from multiple systems into data warehouses or data lakes [10]. Sharing these data beyond organizational boundaries is not viable because anonymization may be impossible for certain data types such as genomic data, and because linking datasets increases re-identification risk. Alternatively, research communities build domain-specific data infrastructures [11], for example in bioinformatics, cohort studies, clinical research, or biobanks. However, the problem of accessing data outside the network remains, and since data are collected for specific uses and duplicated outside the primary data source, this limits record linkage and integration of multimodal data.

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

Another critical aspect is social and cultural. The sensitive nature of health data makes individuals and institutions hesitant to share. From the public perspective, people are more willing to accept and participate in data sharing when informed about existing safeguards and governance mechanisms. They are willing to contribute to science for better care and well-being, but want to decide who can use their data, for what purposes, and ensure data users are accountable for their actions. A survey of 603 secondary data users showed that 56% of researchers willing to share data demand context with access control and want a say or at least knowledge regarding data use [15]. Current data sharing practices do not allow owners to decide who accesses data and for which purposes. Although data sharing and licensing agreements set terms and conditions—such as limitations to specific research purposes, conditions of data transfer, prohibitions on establishing individual identities, or maximum time before data must be destroyed—once data leaves institutional boundaries, no mechanism enforces these policies.

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

FAIR principles become relevant not only for data but also for analytics tasks. In the fragmented data landscape, interoperability and accessibility can be ensured by applying FAIR principles to analytics tasks and system components that interact with these tasks. This paper demonstrates 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 infrastructure to support distributed and federated solutions that utilize data at their original location. Health data are typically produced by diverse sources including care institutions, biomedical researchers, imaging facilities, clinical and population studies, genomic sequencing centers, and citizens themselves. The PHT creates an open ecosystem by making self-contained, machine-readable analytics tasks exchangeable and executable across diverse systems. The PHT does not prescribe specific standards or technologies for data; 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 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 for reusing 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 giving data owners authority to decide and monitor data use. This will eventually lead to creating an Internet of FAIR data and services that operates on personal health data that can never be completely open.

An example application is training a patient survival prediction model, which requires assessing and analyzing large amounts of real-world, high-dimensional, multimodal personal data. In the health domain, this corresponds to information such as longitudinal medical records, diagnostic tests like imaging, genomic profiles, and patient-generated health data and outcomes via apps and wearable devices. To discover hidden patterns, the full dataset should be available to the machine learning task, but privacy-driven data minimization requirements limit personal data to 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 healthcare providers through the PHT infrastructure without data ever leaving organizational boundaries [16, 17].

The PHT defines 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 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 execute a data analytics algorithm to derive insights. They formulate queries and specify the analytics algorithm. The set of all artifacts required to execute the distributed algorithm and return results is called a "Train." A Train is identified by a Digital Persistent Identifier (PID) and contains a self-sufficient message with all information required to transfer code and results between 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: 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 Train reference to the handling Station. Trains are registered in a Train Registry to make them identifiable. The consumer has no direct access to data sources, and humans are entirely decoupled from the computation phase until the algorithm finishes.

Handler (Track): Acts as a gateway between the consumer and curators. It orchestrates communication by receiving self-sufficient Trains from the consumer and forwarding them to selected Stations. It may act as a broker and aggregate results from multiple curators. It manages Train and Station states and logs 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 Track evaluates and aggregates results from different Stations and sends them back to the consumer.

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

---

3. Following FAIR Principles for Distributed Analytics

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

distributed data with distributed analytics.

Clearly, making data self-explanatory and discoverable significantly ensures reusability. However, this may not always be possible, specifically when data are sensitive and not collected for research purposes. For data collected during routine healthcare, for example, it is likely that data are stored in heterogeneous systems following data standards imposed by daily transaction requirements, such as HL7 or DICOM, which might not support desired metadata levels and persistent identification schemes. Therefore, the PHT needs to interact with data repositories that may or may not follow FAIR principles, despite FAIR data being highly desirable. Participating data repositories independently decide their degree of FAIR support. They act as FAIR data points [21] by implementing custom interfaces that support computational tasks reusing data.

The PHT sets machine readability at its core, aiming for maximal interoperability between diverse systems, and is therefore well aligned with FAIR principles. PHT infrastructure components support FAIR principles to varying degrees (Figure 2 [Figure 2: see original paper]).

Station Private: Access to health data has restrictions deriving from original patient consent or institutional data protection policies [22]. These data should be kept in a secure part of the Station inaccessible to external data consumers. FAIR is not a requirement for the private part of data repositories where data and metadata reside, which may follow preferred institutional standards. However, analytics task access should be supported by having queryable consent, a mechanism to link datasets, and a virtual layer for integrated queries over diverse datasets, therefore requiring 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 communication protocols to discover and receive Trains. Analytics tasks can be delivered with open, free, and universally implementable protocols mandating authentication and authorization procedures. Access control to data resides in Station sovereignty, 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 metadata about data repositories and 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 registered in a Train Registry and searchable by their metadata. Train objects are persistently stored in repositories containing all source and environment information required to 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 content and provenance of the analytics task—such as intended use, developer, consent requirements —and task-specific requirements like dependencies, prescribed data standards, and computational resources. They are self-contained, enabling virtualization to support interoperability during execution. Trains are Reusable, designed for use in multiple locations. Licenses and certifications can be assigned to Trains, and they maintain detailed provenance metadata including execution history.

What FAIRness means for a PHT Station: (F) As a data owner, I want to provide enough metadata to be discoverable and published by the 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 providing permission; (I+A) As a Station administrator/dispatcher, I want to set a mechanism to prevent high demand (e.g., prevent Station crashes); (I+A) As a Station administrator, I want to interact with Trains through defined interfaces for providing data input and executing 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 (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 I am sending over public networks is securely transferred (is there a mechanism, e.g., public/private keys); (A) As a data consumer/owner, I want to apply authorization and authentication policies to Train repositories for identity management.

What FAIRness means for PHT Tracks: Train repositories are the building blocks to achieve FAIRness of analytics tasks. They should adopt and follow recommendations for data repositories: persistent identification, application programming interface (API), Train curation and moderation workflows, accessibility, license for reuse, and sustainability [24]. The first recommendation is assigning persistent and global identifiers to each Train. Various identifier schemas such as URIs or DOIs can be employed. Trains deposited in private registries should be described with rich descriptive and operational metadata and can be registered in public repositories such as DataCite. Trains should receive a PID ideally at the earliest workflow state, and to support later operations, the PID should be embedded in the object [25].

Identification of Trains with PIDs and associated machine-readable metadata can facilitate distribution of Trains in a Digital Object Architecture [26]. The second recommendation for FAIR Train repositories is offering 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 in defining and moderating their Trains composed of analytics tasks and meta-

data. Similar to data curation experts, data scientists require tools where they can check, verify, and approve content. Train repository accessibility 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, various options exist for licensing data and database rights [27]. Further investigation should associate licenses with Trains reflecting intellectual property and copyrights of analytics tasks. The final 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 aggregation tasks when required. All communication is logged by the Handler, improving transparency and accountability among involved partners. Table 1 summarizes the FAIR principles supported by the PHT.

---

4. Conclusion and Outlook

The PHT is a novel approach establishing a FAIR distributed data analytics infrastructure enabling (re)use of distributed healthcare data while data owners retain control of their own data. In summary, the PHT: (i) empowers citizens and organizations to control data use in their own repositories for individual and societal benefit; (ii) improves health data usability by lowering data protection barriers while ensuring privacy and confidentiality preservation; (iii) ensures data sovereignty beyond security and privacy by supporting responsible use; (iv) builds trust between data consumers and owners by making analytics processes repeatable, transparent, and auditable; and (v) applies FAIR principles to protocols governing how data analytics interacts with FAIR data points by making analytics tasks themselves FAIR and placing machine readability at its core.

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

The PHT has already been 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 specific diseases at a PHT data Station. The PCC can work with different data sources having different representations (e.g., FHIR, RDF, OMOP-OHDSI, CDISC-ODM) and is agnostic to underlying data. The current implementation works with two data sources: one with RDF based on the Radiation Oncology Ontology, and one Station using FHIR. Other applications include the Varian Learning Portal by Varian

Medical Systems and the open-source software ppDLI by IKNL, both example implementations of distributed learning PHT infrastructures in healthcare.

One use case demonstration involves developing a distributed Bayesian network model to predict dyspnea after radiotherapy for lung cancer patients, developed and used in the Varian Learning portal with data from five different hospitals. The ppDLI implementation currently provides a ready-to-use implementation of the distributed Cox Proportion Hazards algorithm [16]. The 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 that need to process data but cannot share them due to sensitivity, such as the agricultural sector and 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 works on PHT infrastructure development and implementations. O. Kohlbacher (oliver.kohlbacher@uni-tuebingen.de) reviewed the paper 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 PHT development. R. Karim (rezaul.karim@fit.fraunhofer.de) reviewed the manuscript and works 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 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, Switzerland: 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/howdoesdatashield
[14] DataSHIELD –New directions and dimensions. Available at: https://datascience.codata.org/articles/10.533 2017-021/.
[15] What drives academic data sharing? Available at: https://journals.plos.org/plosone/article?id=10.1371/jou
[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.

Note: Figure translations are in progress. See original paper for figures.