Bioinformatics , 2023, pp. 1–7

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

FAIR Data Cube, a FAIR data infrastructure for integrated multi-omics data analysis

Xiaofeng Liao,[1,*] Anna Niehues,[1,2] Casper de Visser,[1] Junda Huang,[1] Cenna Doornbos,[1] Thomas H.A. Ederveen,[1] Purva Kulkarni,[1,2,3] K. Joeri van der Velde,[4] Morris A. Swertz,[4] Martin Brandt,[5] Alain J. van Gool[2,3] and Peter A.C. 't Hoen[1, *]

[1]Medical BioSciences department, Radboud university medical center, Nijmegen, The Netherlands, [2]Translational Metabolic Laboratory, Department of Laboratory Medicine, Radboud university medical center, Nijmegen, The Netherlands, [3]Department of Human Genetics, Radboud university medical center, Nijmegen, The Netherlands, [4]Genomics Coordination Center, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands and [5]SURF, Science Park 140, 1098 XG, Amsterdam, The Netherlands
[*]Corresponding author. XiaoFeng.Liao@radboudumc.nl, Peter-Bram.tHoen@radboudumc.nl

Abstract

Motivation: We are witnessing an enormous growth in the amount of molecular profiling (-omics) data. The integration of multi-omics data is challenging. Moreover, human multi-omics data may be privacy-sensitive and misused to de-anonymize and (re-)identify individuals. Hence, most data is kept in secure and protected silos. Therefore, it remains a challenge to reuse these data without infringing the privacy of the individuals from which the data were derived. Federated analysis of FAIR data is a privacy-preserving solution to make optimal use of these multi-omics data and transform them into actionable knowledge.
Results: The Netherlands X-omics Initiative is a National Roadmap Large-Scale Research Infrastructure aiming for efficient integration of data generated within X-omics and external datasets. To facilitate this, we developed the FAIR Data Cube (FDCube), which adopts and applies the FAIR principles and helps researchers to create FAIR data and metadata, facilitate reuse of their data, and make their data analysis workflows transparent. The FDCube also meets security-by-design and privacy-by-design principles.
Availability: `https://github.com/Xomics/FAIRDataCube`
Contact: Xiaofeng.Liao@radboudumc.nl, Peter-Bram.tHoen@radboudumc.nl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Key words: FAIR, Multi-omics, Data Sovereignty, FAIR Data Cube, Metadata, Federated Analysis

Introduction

It is now widely acknowledged that understanding the mechanisms underlying health and disease requires the concerted study of different molecular levels (DNA, RNA, proteins, metabolites). Moreover, a transition from static simplified views to dynamic comprehensive views on molecular pathways (encompassing (*e.g.* genomics, proteomics and metabolomics) is needed. Currently, this is not simple nor scalable. There is an increasing need to combine -omics data from different sources, but the data and their associated metadata are not always findable, accessible, interoperable, and reusable (FAIR) [31]. For that reason, the Netherlands X-omics Initiative has developed a multi-omics data infrastructure that facilitates FAIR-compliant multi-omics data storage and analysis. The proposed data infrastructure provides an analysis environment for (federated) data handling and analysis meeting the security-by-design and privacy-by-design principles.

This paper introduces our solution of integrated analysis on FAIR multi-omics data in decentralized databases. In the remainder of this paper, section 2 investigates existing work in this research direction. Section 3 presents the design and implementation of the FAIR Data Cube (FDCube) and section 4 showcases the use of FDCube in the Trusted World of Corona project[11]. Finally, section 5 discusses further developments.

Related work

There are several tools that aid researchers in managing research metadata in a FAIR manner, for instance the FAIR Data Station[23], the FAIR-in-a-box[5] approach, and the DataFAIRifier[1]. Most of these tools focus on the production of FAIR data, including ingestion, generation, and publication.

For a more comprehensive coverage of FAIR processes including data management, data security, data exchange, and

federated analysis, additional tools are required. For example, MOLGENIS is an open-source web-application covering the typical flow of human genomics data including data collection, management, analysis, visualization, and sharing, as well as offering support to make data FAIR[30]. MOLGENIS can be hosted on-site and stores the data locally in a PostgreSQL database. This offers all the advantages of a true database including a local access control system (in light of the European General Data Protection Regulation) with detailed data management.

The Personal Health Train (PHT)[14] concept is underlying a number of approaches for decentralised analysis of health-related data. The essence of the PHT approach is the analogy of a station representing the data source and a train representing the research question (or a computational request) visiting the data stations. Stations range from very large databases to small personal lockers containing the data of one person. Each station has its own set of house rules describing what a visiting 'train' is allowed to do with its data[14]. By moving trains towards stations rather than moving data, copying of data is avoided, data remains under complete control of the person or institute generating the data, and privacy concerns around data sharing are alleviated.

DataSHIELD[18]implements the idea of bringing algorithms to the data to ensure data privacy and security. DataSHIELD facilitates (co-)analysis of (harmonised) biomedical, healthcare and social-science data stored at one or multiple locations. The analysis requests are sent from a central analysis machine to several data-holding machines, which store the harmonised data to be co-analysed. The datasets are then analysed simultaneously, but in parallel. MOLGENIS developed a DataSHIELD implementation called Armadillo in its MOLGENIS suite.

Vantage6[22, 28] is a different implementation of the PHT concept. Vantage6 enables collaboration between multiple parties to participate in one or multiple studies across multiple data stations.

In terms of programming language, DataSHIELD restricts itself to a single language (R)[24] and to a pre-defined library of functions and algorithms. By contrast, using Vantage6, the researcher can pose a request to use their preferred programming language, as long as the language is supported by the targeted data station.

To advance and further build upon the currently available federated, FAIR solutions for the scientific community, we here present the FAIR Data Cube (FDCube) for public use under an open MIT license. In contrast to the more generic MOLGENIS Armadillo approach, FDCube contains specialised services for the analysis of multi-omics data. The FDCube is developed based on the principle that data should be "as open as possible and as closed as necessary" [17]. By incorporating a FAIR Data Point (FDP) component, the metadata can be as open as possible and made FAIR-at-the-source. By integrating a Vantage6 component, the data security/privacy can be ensured by collaborated federated analysis.

Result

The FDCube is a technological framework for the storage, analysis and integration of multi-omics data. The FDcube reuses and extends existing open software components/modules and initiatives. This includes the FAIR Data Point[16] and Vantage6[22]. Further elements of the FDCube are the Investigation-Study-Assay (ISA) metadata framework[27, 20] for capturing general study metadata, sample (including basic sample characteristics), and assay metadata, and the Phenopackets[21] standards for capturing phenotypic description of a patient/sample. The concept of the FDCube is illustrated in Fig 1 and detailed below from the perspective of a dataset owner and a researcher as a user of that dataset, respectively. The complete and detailed documentation on the FDCube can also be found at `https://github.com/Xomics/FAIRDataCube/wiki`.

Dataset owner

A dataset owner registers their dataset by publishing the metadata on a FAIR Data Point (FDP). The FDP is a metadata repository that provides public access to metadata in accordance with the FAIR principles[16]. The FDP helps dataset owners to publish the metadata of their dataset, and facilitates researchers (dataset users) to find and access information (metadata) about the registered datasets, including pointers to that data (irrespective of data access restrictions and licenses, which is typically arranged at the location of the data store/source).

Considering the various metadata formats adopted by the different X-omics communities, it is reasonable to adopt a standard metadata format as a template for submitting the metadata. To this purpose, we employed the Investigation-Study-Assay (ISA) metadata framework[27, 20] as our basic framework to capture and standardize study (design) information from the different -omics metadata schemes. The ISA metadata schema is commonly adopted by the research community for submission of metabolomics data, for example by EMBL-EBI's MetaboLights repository[8].

In biomedical studies, clinical characteristics and phenotype information of the study subjects may be collected in addition to (-omics or other) measurements data. This information is essential for making interpretations from research experimental data. Thus, phenotype data need to be standardized as well, so that researchers and clinicians can more easily link phenotypes to experimental data. To achieve this, the Phenopackets framework[21] developed by Global Alliance for Genomics and Health (GA4GH) was adopted. This framework comprises a comprehensive data structure (model), using common ontology terms, to categorise and connect different types of phenotype data.

Researcher

The researcher can be both a data set owner and a data set consumer. As a dataset consumer, the research can search a FDP, which is part of a FDCube, for any dataset of interest. Since all metadata is represented in a linked data format, the researcher can conduct semantic searches on datasets and their corresponding study information by using the SPARQL Protocol and RDF Query Language (SPARQL) query interface. The information that can be queried is the ontologized description of, for instance: samples and their (biological) source; sample preparation; methods and techniques applied; (-omics) measurement and (data) analysis strategies, workflows and reports, including the detected (molecular) data features; research group affiliations. Example questions that may be asked are:

1. Find all studies which use mass spectrometry-based metabolomics and study a specific metabolic disorder;

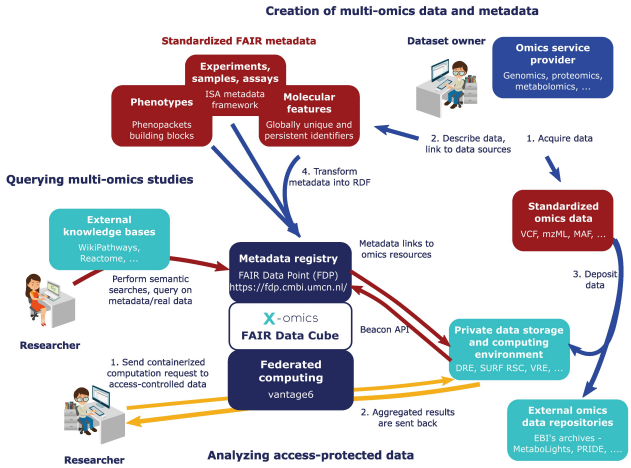

Fig. 1. The concept of the FDCube. Dataset owners and researchers as a dataset consumer can both benefit from FDCube on various aspects, including creation of multi-omics data and metadata, querying multi-omics studies and analyzing access-protected data via federated analysis.

2. Find datasets with more than two -omics types and more than a 100 individuals;
3. Find measurements for proteins and metabolites that belong to a particular metabolic pathway.

To explore more complex research questions, the researcher could raise a computational request to the dataset owner. This is achieved by the Vantage6 component of the FDCube.

Demonstration of FDCube in TWOC

We adopted the Trusted World of Corona (TWOC) project to demonstrate how to utilize the FDCube for integrated multi-omics federated analysis. The TWOC project aims to contribute to a more sustainable, innovative high-quality and person-oriented healthcare system. To this end, they created a platform in which humans and machines can meet based on FAIR data, protocols and algorithms.

In Fig 2, we provide an example of the creation and application of the FDCube based on a public dataset on COVID-19 featuring multi-omics patient data by Su et al., 2020[29], which was FAIRified as part of the TWOC project.

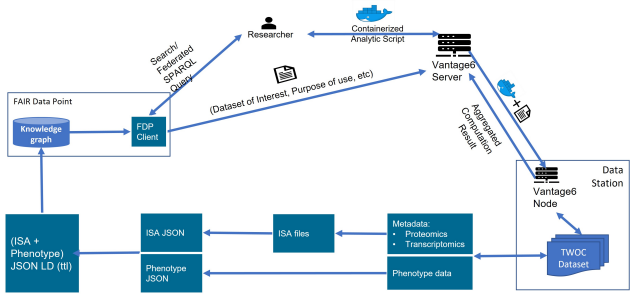

Fig. 2. Use of FDCube in the TWOC demonstrator. The FDCube workflows covers various functions including creating and publishing metadata, browsing and querying the metadata on FDP, and creating and running federated data analysis.

Below is an overview of the workflows for creating, filling, and using the FDCube.

Storage of raw and processed omics data

A multi-modal dataset from COVID-19 patients[29] was prepared, harmonized and FAIRified as part of the TWOC project. The dataset consists of paired omics data layers describing transcriptomics, proteomics and metabolomics of blood samples, and includes comprehensive phenotype information. The dataset is publicly accessible at TWOC's GitHub repository[12].

To allow interactive and joint querying of data and metadata, we store the processed -omics data along with their feature annotation files. These are both stored in a flat-text tabular *.csv* format, with features as rows and samples as columns.

Creation of metadata

In the TWOC project, both the ISA metadata schema and Phenopackets schema are adopted. The ISA metadata schema is used as a standard metadata schema to capture metadata about (-omics) experiments, and serializes in an ISA-json file using ISA tools[26, 20]. The ISA tools API provides additional functionalities to convert the ISA objects into linked data.

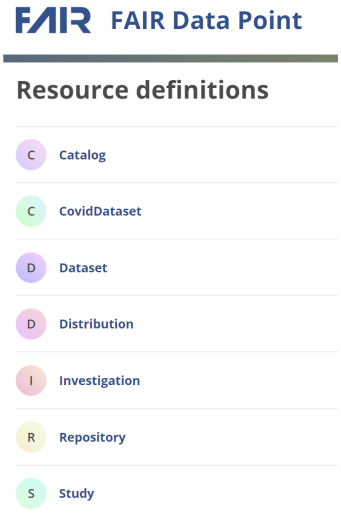

Fig. 3. Overview of all FDP resources. The Investigation and Study resources are defined in addition to the FDP's default resources.

The FAIR Data Point adopts the W3C's Data Catalog Vocabulary (DCAT)[15] as its basic metadata schema to capture generic information of the registered datasets and their distributions. To host the experimental metadata in the ISA schema, we defined extra resources and Shapes Constraint Language (SHACL) shapes for the investigation and study. Fig 3 shows all the FDP resources, including the additional investigation and study files. A detailed SHACL shape of the new resources can be found on the FDCube GitHub repository[10].

Example scripts[13] are provided to assist researchers in using these frameworks to capture study and experimental (meta)data as well as phenotype information and to share it on a FDP server.

The Investigation and Study part of ISA is made DCAT-compatible and is used to create an input form to publish metadata on the FDP. Given the flexibility in the Assay part, a potential solution is to first import the ISA metadata and the accompanying phenotype metadata into the triplestore behind the FDP, like GraphDB or Blazegraph. After that, a subset of the metadata can be selected and publically displayed for browsing. The selection of a triplestore is an option that can be selected by the FDCube user. A containerized environment to utilize the ISA-API[7], coupled with the ISA cookbook [6], was created for researchers to FAIRify experimental metadata that is used as input for the FDCube. Moreover, we developed a containerized workflow for the automatic submission of the TWOC clinical metadata to FDP[3].

For phenotype data, a python script[13] was developed based on the phenopackets data schema to automatically convert unFAIRified phenotype information into csv format. We then wrote a YARRRML[19] template that embedded the phenopackets RDF-schema [9], making use of the transformation service in FAIR-in-a-box[5]. This converts the csv file into linked data.

.

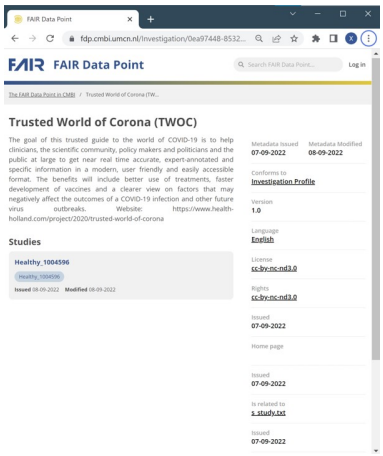

Fig. 4. FAIRified metadata of TWOC dataset published on FAIR Data Point portal

Querying of metadata

The FAIR Data Point can display complete/partial metadata in a human-readable portal for browsing, searching and querying. The FAIRified metadata of the TWOC dataset was published on a FDP portal [4] as shown in Fig 4. A SPARQL query can be run against the metadata via the a query portal to gain deeper knowledge of a dataset, as illustrated in Fig 5. The FDP portal provides a user interface where users can design SPARQL queries. After finding an interesting dataset via browsing or by SPARQL, the researcher could further run follow-up analyses on the target dataset by raising a computation request to the Vantage6 server and retrieve the returning results from the data station via Vantage6.

Running a data analysis script

Vantage6 delivers the user's computational request to a data station. A computation request consists of:

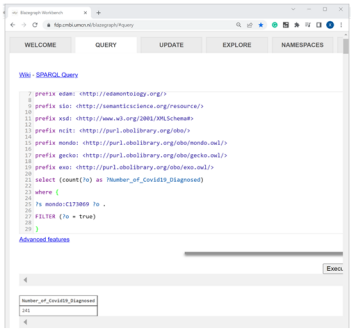

Fig. 5. The SPARQL query portal.

- A reference to a Docker image, which contains the code (computation) that the researcher would like to run on the target dataset;
- A list describing the dataset of interest and its purpose-of-use.

The Vantage6 server handles authentication, keeps track of all computation requests, assigns them to nodes for computation, and stores the returning results of the analyses. The Vantage6 server could also host a private Docker registry.

A Vantage6 node is typically installed at a dataset station. For security reason, the dataset station could stay in an access-protected environment, for example, in a Digital Research Environment (DRE)[2], which is a cloud based, globally available research environment.

Fig 6 shows the Vantage6 user interface at which a researcher can create a task.

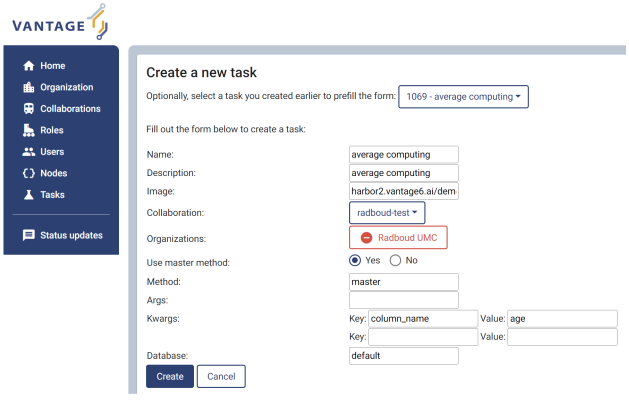

Fig. 6. Create a task in the Vantage6 user interface

In this example, we used an averaging algorithm hosted on Docker Hub[1]. This algorithm expects an argument *column_name* to be defined, and will compute the average over that column. We specified in the *kwargs* fields the parameter 'column_name' with value 'age'. The averaging algorithm is dispatched to run on a Vantage6 node, where the dataset is stored. In this example, the dataset is a .csv file prepared from TWOC, which contains a column titled *age*. The *Database* field in Fig 6 is labeled *default*, which is configurable in the

[1] `harbor2.vantage6.ai/demo/average`

Vantage6 node configuration file. For simplicity, this task is created for a collaboration with only one organization (in our example Radboudumc).

Fig 7 shows the result of running the averaging algorithm on the patients' age in the TWOC dataset, which specifically calculates the average value in the column labelled *age*. This result can be passed back as the response to the computation request.

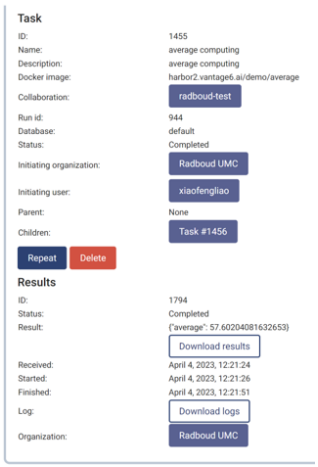

Fig. 7. Vantage 6 task running result

Conclusion

We have created the FAIR Data Cube, a software and programmatic infrastructure to make -omics data FAIR, and to facilitate the management, reuse, integration and analysis of biomedical (-omics) data, while ensuring data sovereignty, by utilizing Vantage6' capability of 'bringing research questions to data' rather than 'sending data to research questions'. Vantage6's management capability covers comprehensive aspects (including organization, collaboration, users, roles, nodes and tasks), and makes FDCube a useful platform to carry out cross-organization federated analysis on decentralized datasets.

We used the FDCube in the TWOC project to demonstrate its capability and usage on, creating and publishing ISA and phenotype meta data, browsing and querying the metadata on FDP, and creating and running federated data analysis on a real dataset.

There are several ways to improve and extend the design and implementation of the current FDCube. For example, a Beacon[25] component can be integrated into FDCube. The reason for this integration is that a FDP (by design) only exposes metadata of datasets. In contrast, Beacon allows for more insights about the presence/absence of specific genomic mutation in a set of data[25]. The combined information from both metadata (via FDP) and real data (via Beacon query), would help a researcher to get more insights into possibly available datasets before designing a data analysis request as dictated by the researcher's study questions.

Another potential work would be to integrate DataSHIELD and Vantage6 to grant users of Vantage6 access to rich analysis algorithms in DataSHIELD.

Competing interests

No competing interest is declared.

Author contributions statement

P.A.C.H., A.J.G, M.A.S conceived the project. J.H. worked on phenotype data modelling. A.N., C.V worked on ISA metadata. T.E. managed connection to the TWOC project and FAIRification of the presented dataset. P.K worked on lipidomics metadata. C.D. promoted FDCube and provided scientific feedback. M.B supported the hosting environment. K.J.V provided insights from MOLGENIS perspective. A.N. presented the high level concept diagram. X.L. implemented and set up the architecture with help from all team members. X.L. wrote the manuscript with critical input and revisions from A.N., C.D., C.V., J.H., T.E., P.A.C.H, P.K., K.J.V., A.J.G. All authors reviewed the manuscript.

Acknowledgments

This work was funded by a Dutch Research Council (NWO) grant to The Netherlands X-omics Initiative (project 184.034.019), a Horizon2020 grant to the European Joint Programme on Rare Diseases (grant agreement Number 82557), a Horizon2020 grant to the EATRIS-Plus project (grant agreement Number 871096), and a LSH HealthHolland grant to the Trusted World of Corona (TWOC) consortium.

References

1. Datafairifier. `https://github.com/MaastrichtU-CDS/DataFAIRifier`. Accessed: 2020-04-19.
2. Digital research environment. `https://www.radboudumc.nl/en/research/radboud-technology-centers/data-stewardship/digital-research-environment`. Accessed: 2020-04-19.
3. Docker container to run the fdp submission script. `https://github.com/Xomics/FAIRDataCube/tree/master/Docker`. Accessed: 2020-04-19.
4. The fair data point in cmbi. `https://fdp.cmbi.umcn.nl`. Accessed: 2020-04-19.
5. Fiab: Fair-in-a-box. `https://github.com/ejp-rd-vp/FiaB`. Accessed: 2020-04-19.
6. Isa tools api. `https://isa-tools.org/isa-api/content/index.html`. Accessed: 2020-04-19.
7. Isa tools environment. `https://github.com/Xomics/Isatools_environment`. Accessed: 2020-04-19.
8. Metabolights. `https://www.ebi.ac.uk/metabolights/`. Accessed: 2020-04-19.
9. Phenopackets rdf sschema. `https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema`. Accessed: 2020-04-19.
10. Setting up an isa compatible fair data point instance. `https://github.com/Xomics/FAIRDataCube/blob/master/ISA_FDPSettingup.md`. Accessed: 2020-04-19.
11. Trust world of corona. `https://www.health-holland.com/project/2020/trusted-world-of-corona`. Accessed: 2020-04-19.
12. Twoc demonstrator. `https://github.com/Xomics/TWOCdemonstrator/tree/main/data/Su_2020_original/phenotypes_in_modules`. Accessed: 2020-04-19.

13. Twoc demonstrator tools. `https://github.com/Xomics/TWOCdemonstrator/tree/main/tools`. Accessed: 2020-04-19.
14. Oya Beyan, Ananya Choudhury, Johan van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Md. Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino da Silva Santos, and Andre Dekker. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. *Data Intelligence*, 2(1-2):96–107, 01 2020.
15. World Wide Web Consortium et al. Data catalog vocabulary (dcat). 2014.
16. Luiz Olavo Bonino da Silva Santos, Kees Burger, Rajaram Kaliyaperumal, and Mark D. Wilkinson. FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. *Data Intelligence*, pages 1–21, 09 2022.
17. European Commission. Directorate-General for Research Innovation. H2020 programme guidelines on fair data management in horizon 2020. 2016.
18. Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, Rebecca Wilson, Oliver Butters, Barnaby Murtagh, Ipek Demir, Dany Doiron, Lisette Giepmans, Susan E Wallace, Isabelle Budin-Ljøsne, Carsten Oliver Schmidt, Paolo Boffetta, Mathieu Boniol, Maria Bota, Kim W Carter, Nick deKlerk, Chris Dibben, Richard W Francis, Tero Hiekkalinna, Kristian Hveem, Kirsti Kvaløy, Sean Millar, Ivan J Perry, Annette Peters, Catherine M Phillips, Frank Popham, Gillian Raab, Eva Reischl, Nuala Sheehan, Melanie Waldenberger, Markus Perola, Edwin van den Heuvel, John Macleod, Bartha M Knoppers, Ronald P Stolk, Isabel Fortier, Jennifer R Harris, Bruce HR Woffenbuttel, Madeleine J Murtagh, Vincent Ferretti, and Paul R Burton. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*, 43(6):1929–1944, 09 2014.
19. Pieter Heyvaert, Ben De Meester, Anastasia Dimou, and Ruben Verborgh. Declarative rules for linked data generation at your fingertips! In Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam, editors, *The Semantic Web: ESWC 2018 Satellite Events*, pages 213–217, Cham, 2018. Springer International Publishing.
20. David Johnson, Dominique Batista, Keeva Cochrane, Robert P Davey, Anthony Etuk, Alejandra Gonzalez-Beltran, Kenneth Haug, Massimiliano Izzo, Martin Larralde, Thomas N Lawson, Alice Minotto, Pablo Moreno, Venkata Chandrasekhar Nainala, Claire O'Donovan, Luca Pireddu, Pierrick Roger, Felix Shaw, Christoph Steinbeck, Ralf J M Weber, Susanna-Assunta Sansone, and Philippe Rocca-Serra. ISA API: An open platform for interoperable life science experimental metadata. *GigaScience*, 10(9), 09 2021. giab060.
21. Markus S. Ladewig, Julius O. B. Jacobsen, Alex H. Wagner, Daniel Danis, Baha El Kassaby, Michael Gargano, Tudor Groza, Michael Baudis, Robin Steinhaus, Dominik Seelow, Nikolaos E. Bechrakis, Christopher J. Mungall, Paul N. Schofield, Olivier Elemento, Lindsay Smith, Julie A. McMurry, Monica Munoz-Torres, Melissa A. Haendel, and Peter N. Robinson. Ga4gh phenopackets: A practical introduction. *Advanced Genetics*, n/a(n/a):2200016.
22. Arturo Moncada-Torres, Frank Martin, Melle Sieswerda, Johan van Soest, and Gijs Geleijnse. Vantage6: an open source privacy preserving federated learning infrastructure for secure insight exchange. In *AMIA Annual Symposium Proceedings*, pages 870–877, 2020.
23. Bart Nijsse, Peter J. Schaap, and Jasper J. Koehorst. Fair data station for lightweight metadata management & validation of omics studies. *bioRxiv*, 2022.
24. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.
25. Jordi Rambla, Michael Baudis, Roberto Ariosa, Tim Beck, Lauren A. Fromont, Arcadi Navarro, Rahel Paloots, Manuel Rueda, Gary Saunders, Babita Singh, John D. Spalding, Juha Törnroos, Claudia Vasallo, Colin D. Veal, and Anthony J. Brookes. Beacon v2 and beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. *Human Mutation*, 43(6):791–799, 2022.
26. Philippe Rocca-Serra, Eamonn Maguire, Chris Taylor, Dawn Field, Timo Wittenberger, Annapaola Santarsiero, Alejandra Gonzalez-Beltran, and Susanna-Assunta Sansone. 7 - investigation-study-assay, a toolkit for standardizing data capture and sharing. In Lee Harland and Mark Forster, editors, *Open Source Software in Life Science Research*, Woodhead Publishing Series in Biomedicine, pages 173–188. Woodhead Publishing, 2012.
27. S.A. Sansone, P. Rocca Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C.T.A. Evelo, M.J. Forster, P. Gaudet, J. Gilbert, C. Goble, J.L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S.J. Ho Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C.E. Shamu, C.A. Shang, C. Steinbeck, A. Trefethen, B. Jones, K. Wolstencroft, I. Xenarios, and W. Hide. Toward interoperable bioscience data. *Nature Genetics*, 44(2):121–126, February 2012.
28. Djura Smits, Bart van Beusekom, Frank Martin, Lourens Veen, Gijs Geleijnse, and Arturo Moncada-Torres. An improved infrastructure for privacy-preserving analysis of patient data. In *Proceedings of the International Conference of Informatics, Management, and Technology in Healthcare (ICIMTH)*, volume 295, pages 144–147, 2022.
29. Yapeng Su, Daniel Chen, Dan Yuan, Christopher Lausted, Jongchan Choi, Chengzhen L. Dai, Valentin Voillet, Venkata R. Duvvuri, Kelsey Scherler, Pamela Troisch, Priyanka Baloni, Guangrong Qin, Brett Smith, Sergey A. Kornilov, Clifford Rostomily, Alex Xu, Jing Li, Shen Dong, Alissa Rothchild, Jing Zhou, Kim Murray, Rick Edmark, Sunga Hong, John E. Heath, John Earls, Rongyu Zhang, Jingyi Xie, Sarah Li, Ryan Roper, Lesley Jones, Yong Zhou, Lee Rowen, Rachel Liu, Sean Mackay, D. Shane O'Mahony, Christopher R. Dale, Julie A. Wallick, Heather A. Algren, Michael A. Zager, Wei Wei, Nathan D. Price, Sui Huang, Naeha Subramanian, Kai Wang, Andrew T. Magis, Jenn J. Hadlock, Leroy Hood, Alan Aderem, Jeffrey A. Bluestone, Lewis L. Lanier, Philip D. Greenberg, Raphael Gottardo, Mark M. Davis, Jason D. Goldman, and James R. Heath. Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. *Cell*, 183(6):1479–1495.e20, 2020.
30. K Joeri van der Velde, Floris Imhann, Bart Charbon, Chao Pang, David van Enckevort, Mariska Slofstra,

Ruggero Barbieri, Rudi Alberts, Dennis Hendriksen, Fleur Kelpin, et al. Molgenis research: advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics*, 35(6):1076–1078, 2019.

31. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.