medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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], Thomas H.A. Ederveen[1], Cenna Doornbos[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], 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.
[5]SURF, Science Park 140, 1098 XG, Amsterdam, The Netherlands.

*Corresponding author(s). E-mail(s): XiaoFeng.Liao@radboudumc.nl; Peter-Bram.tHoen@radboudumc.nl;
Contributing authors: annaniehues@eatris.eu; Casper.deVisser@radboudumc.nl; Junda.Huang@radboudumc.nl; Tom.Ederveen@radboudumc.nl; Cenna.Doornbos@radboudumc.nl; Purva.Kulkarni@radboudumc.nl; joeriv@gmail.com; m.a.swertz@gmail.com; martin.brandt@surf.nl; Alain.vanGool@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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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.

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

1 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) [1]. 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[2]. Finally, section 5 discusses further developments.

2 Related work

There are several tools that aid researchers in managing research metadata in a FAIR manner, for instance the FAIR Data Station[3], the FAIR-in-a-box[4] approach, and the DataFAIRifier[5]. 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[6]. MOLGENIS can be

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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)[7] 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[7]. 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[8]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[9, 10] 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)[11] 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" [12]. By incorporating a FAIR Data Point (FDP) component[13], the metadata can be as open as possible and made FAIR-at-the-source. By integrating a Vantage6 component[9], the data security/privacy can be ensured by collaborated federated analysis.

3 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[13] and Vantage6[9]. Further elements of the FDCube are the Investigation-Study-Assay (ISA) metadata framework[14, 15] for capturing general study metadata, sample (including basic sample characteristics), and assay metadata, and the Phenopackets[16] standards for capturing phenotypic description of a patient/sample. The concept of the FDCube

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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.

3.1 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[13]. 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[14, 15] 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[17].

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[16] 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.

3.2 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;
2. Find datasets with more than two -omics types and more than 100 individuals;

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

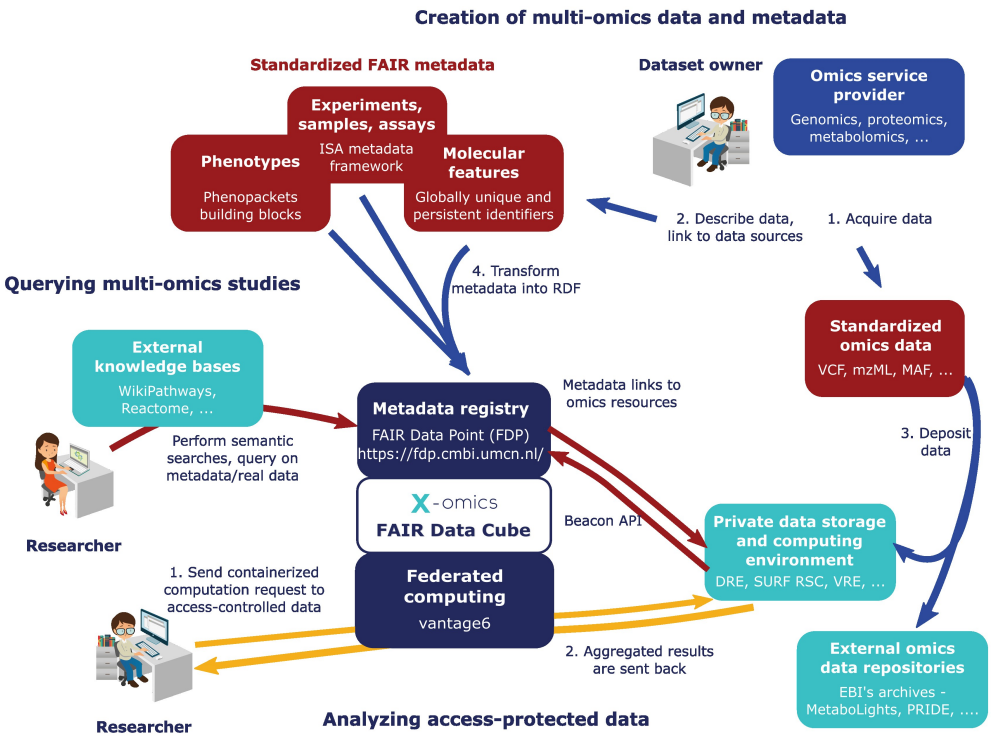

Fig. 1 The concept of the FDCube. Dataset owners and researchers as a dataset consumer can both benefit from FDCube on various aspects, including creation of multi-omics data and metadata, querying multi-omics studies and analyzing access-protected data via federated analysis.

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.

4 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[18], which was FAIRified as part of the TWOC project.

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

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

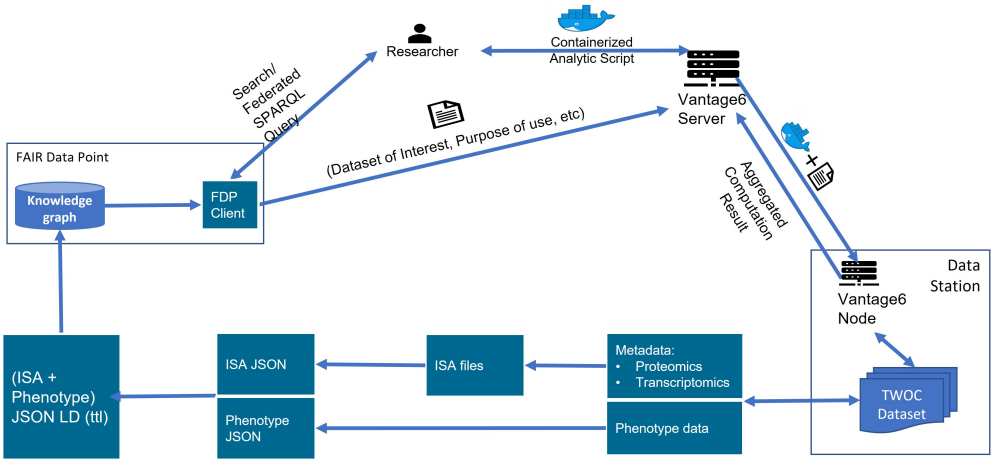

Fig. 2 Use of FDCube in the TWOC demonstrator. The FDCube workflows covers various functions including creating and publishing metadata, browsing and querying the metadata on FDP, and creating and running federated data analysis.

4.1 Storage of raw and processed omics data

A multi-modal dataset from COVID-19 patients[18] 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[19].

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.

4.2 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[15, 20]. The ISA tools also provided additional functionalities to convert the ISA objects into linked data, for example a ttl (Terse RDF Triple Language) [21] file.

Example scripts[22] are provided to assist researchers in capturing study and experimental (meta)data. For phenotype data, a python script[22] was developed based on the phenopackets data schema to automatically convert unFAIRified phenotype information into *csv* format. A YARRRML[23] template was written that embedded the phenopackets RDF-schema [24], making use of the transformation service in FAIR-in-a-box[4]. This converts the csv file into linked data. The final output linked data, including study and experimental (meta)data as well as phenotype information, are uploaded into the triplestore within the FDP. This FAIRified linked data can then be queried using SPARQL to extract valuable insights.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Moreover, a containerized environment to utilize the ISA-API[25], coupled with the ISA cookbook [26], was created to assist researchers in FAIRifying experimental metadata that is used as input for the FDCube.

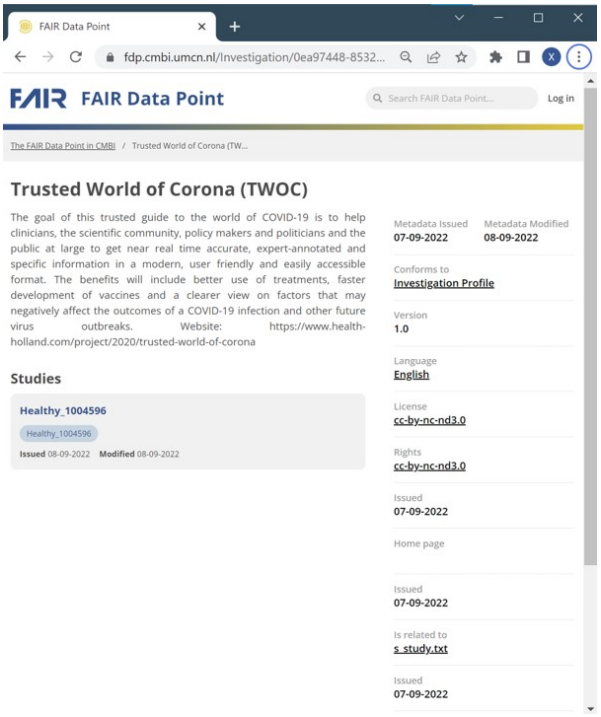

Fig. 3 FAIRified metadata of TWOC dataset published on FAIR Data Point portal

4.3 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 [27] as shown in Fig 3. 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 4. 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.

4.4 Running a data analysis script

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

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

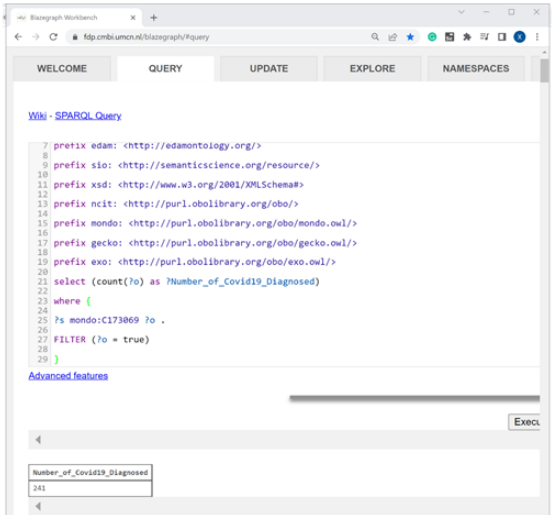

Fig. 4 The SPARQL query portal.

- 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)[28], which is a cloud based, globally available research environment.

Fig 5 shows the Vantage6 user interface at which a researcher can create a task.

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 5 is labeled *default*, which is configurable in the Vantage6 node configuration file. For simplicity, this task is created for a collaboration with only one organization (in our example Radboudumc).

Fig 6 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.

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

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

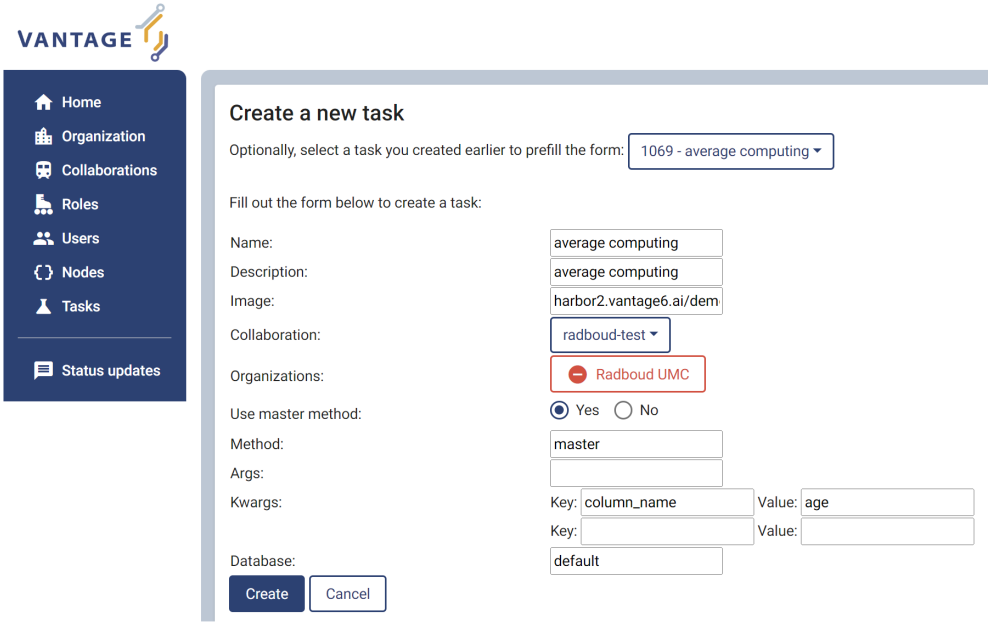

Fig. 5 Create a task in the Vantage6 user interface

5 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[29] 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[29]. 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.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

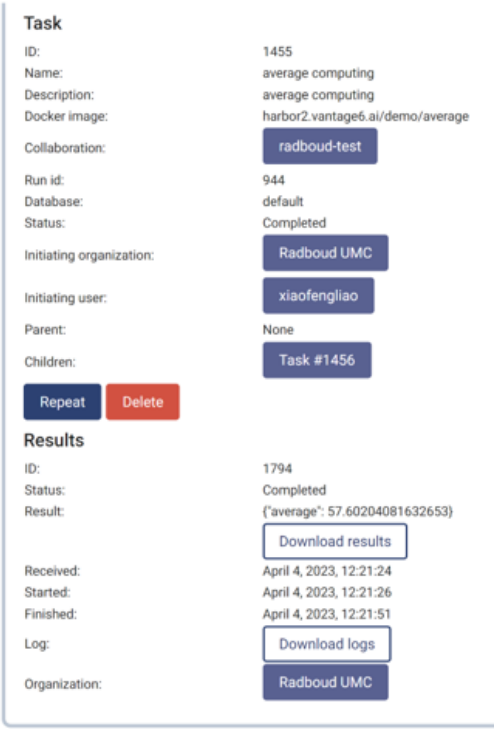

Fig. 6 Vantage 6 task running result

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

Declarations

- Funding
  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 825575), 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.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
  No competing interest is declared.
- Ethics approval
  Not applicable
- Consent to participate
  Not applicable
- Consent for publication
  Not applicable

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

- Availability of data and materials
  Not applicable
- Code availability
  https://github.com/Xomics/FAIRDataCube
- Authors' contributions
  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.

References

[1] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., *et al.*: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)

[2] Trust World of Corona. https://www.health-holland.com/project/2020/trusted-world-of-corona. Accessed: 2020-04-19

[3] Nijsse, B., Schaap, P.J., Koehorst, J.J.: Fair data station for lightweight metadata management & validation of omics studies. bioRxiv (2022) https://doi.org/10.1101/2022.08.03.502622 https://www.biorxiv.org/content/early/2022/08/05/2022.08.03.502622.full.pdf

[4] FiaB: FAIR-in-a-box. https://github.com/ejp-rd-vp/FiaB. Accessed: 2020-04-19

[5] DataFAIRifier. https://github.com/MaastrichtU-CDS/DataFAIRifier. Accessed: 2020-04-19

[6] Velde, K.J., Imhann, F., Charbon, B., Pang, C., Enckevort, D., Slofstra, M., Barbieri, R., Alberts, R., Hendriksen, D., Kelpin, F., *et al.*: Molgenis research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics **35**(6), 1076–1078 (2019)

[7] Beyan, O., Choudhury, A., Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., Karim, M.R., Dumontier, M., Decker, S., Silva Santos, L.O.B., Dekker, A.: Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intelligence **2**(1-2), 96–107 (2020)

[8] Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E.M., Minion, J., Boyd, A.W., Newby, C.J., Nuotio, M.-L., Wilson, R., Butters,

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

O., Murtagh, B., Demir, I., Doiron, D., Giepmans, L., Wallace, S.E., Budin-Ljøsne, I., Oliver Schmidt, C., Boffetta, P., Boniol, M., Bota, M., Carter, K.W., deKlerk, N., Dibben, C., Francis, R.W., Hiekkalinna, T., Hveem, K., Kvaløy, K., Millar, S., Perry, I.J., Peters, A., Phillips, C.M., Popham, F., Raab, G., Reischl, E., Sheehan, N., Waldenberger, M., Perola, M., Heuvel, E., Macleod, J., Knoppers, B.M., Stolk, R.P., Fortier, I., Harris, J.R., Woffenbuttel, B.H., Murtagh, M.J., Ferretti, V., Burton, P.R.: DataSHIELD: taking the analysis to the data, not the data to the analysis. International Journal of Epidemiology **43**(6), 1929–1944 (2014) https://doi.org/10.1093/ije/dyu188 https://academic.oup.com/ije/article-pdf/43/6/1929/18482399/dyu188.pdf

[9] Moncada-Torres, A., Martin, F., Sieswerda, M., Soest, J., Geleijnse, G.: Vantage6: an open source privacy preserving federated learning infrastructure for secure insight exchange. In: AMIA Annual Symposium Proceedings, pp. 870–877 (2020)

[10] Smits, D., Beusekom, B., Martin, F., Veen, L., Geleijnse, G., Moncada-Torres, A.: An improved infrastructure for privacy-preserving analysis of patient data. In: Proceedings of the International Conference of Informatics, Management, and Technology in Healthcare (ICIMTH), vol. 295, pp. 144–147 (2022)

[11] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing. https://www.R-project.org/

[12] Research & Innovation., E.C.D.-G.: H2020 programme guidelines on fair data management in horizon 2020 (2016)

[13] Silva Santos, L.O.B., Burger, K., Kaliyaperumal, R., Wilkinson, M.D.: FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. Data Intelligence, 1–21 (2022) https://doi.org/10.1162/dint_a_00160

[14] Sansone, S.A., Rocca Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.A., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C.T.A., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Ho Sui, S.J., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Jones, B., Wolstencroft, K., Xenarios, I., Hide, W.: Toward interoperable bioscience data. Nature Genetics **44**(2), 121–126 (2012) https://doi.org/10.1038/ng.1054

[15] Johnson, D., Batista, D., Cochrane, K., Davey, R.P., Etuk, A., Gonzalez-Beltran, A., Haug, K., Izzo, M., Larralde, M., Lawson, T.N., Minotto, A., Moreno, P., Nainala, V.C., O'Donovan, C., Pireddu, L., Roger, P., Shaw, F., Steinbeck, C., Weber, R.J.M., Sansone, S.-A., Rocca-Serra, P.: ISA API: An open platform for interoperable life science

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

experimental metadata. GigaScience **10**(9) (2021) https://doi.org/10.1093/gigascience/giab060 https://academic.oup.com/gigascience/article-pdf/10/9/giab060/40394493/giab060_reviewer_3_report_revision_1.pdf. giab060

[16] Ladewig, M.S., Jacobsen, J.O.B., Wagner, A.H., Danis, D., El Kassaby, B., Gargano, M., Groza, T., Baudis, M., Steinhaus, R., Seelow, D., Bechrakis, N.E., Mungall, C.J., Schofield, P.N., Elemento, O., Smith, L., McMurry, J.A., Munoz-Torres, M., Haendel, M.A., Robinson, P.N.: Ga4gh phenopackets: A practical introduction. Advanced Genetics **n/a**(n/a), 2200016 https://doi.org/10.1002/ggn2.202200016 https://onlinelibrary.wiley.com/doi/pdf/10.1002/ggn2.202200016

[17] MetaboLights. https://www.ebi.ac.uk/metabolights/. Accessed: 2020-04-19

[18] Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C.L., Voillet, V., Duvvuri, V.R., Scherler, K., Troisch, P., Baloni, P., Qin, G., Smith, B., Kornilov, S.A., Rostomily, C., Xu, A., Li, J., Dong, S., Rothchild, A., Zhou, J., Murray, K., Edmark, R., Hong, S., Heath, J.E., Earls, J., Zhang, R., Xie, J., Li, S., Roper, R., Jones, L., Zhou, Y., Rowen, L., Liu, R., Mackay, S., O'Mahony, D.S., Dale, C.R., Wallick, J.A., Algren, H.A., Zager, M.A., Wei, W., Price, N.D., Huang, S., Subramanian, N., Wang, K., Magis, A.T., Hadlock, J.J., Hood, L., Aderem, A., Bluestone, J.A., Lanier, L.L., Greenberg, P.D., Gottardo, R., Davis, M.M., Goldman, J.D., Heath, J.R.: Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. Cell **183**(6), 1479–149520 (2020) https://doi.org/10.1016/j.cell.2020.10.037

[19] TWOC demonstrator. https://github.com/Xomics/TWOCdemonstrator/tree/main/data/Su_2020_original/phenotypes_in_modules. Accessed: 2020-04-19

[20] Rocca-Serra, P., Maguire, E., Taylor, C., Field, D., Wittenberger, T., Santarsiero, A., Gonzalez-Beltran, A., Sansone, S.-A.: 7 - investigation-study-assay, a toolkit for standardizing data capture and sharing. In: Harland, L., Forster, M. (eds.) Open Source Software in Life Science Research. Woodhead Publishing Series in Biomedicine, pp. 173–188. Woodhead Publishing, ??? (2012). https://doi.org/10.1533/9781908818249.173 . https://www.sciencedirect.com/science/article/pii/B9781907568978500072

[21] RDF 1.1 Turtle. http://www.w3.org/TR/2014/REC-turtle-20140225/

[22] TWOC Demonstrator Tools. https://github.com/Xomics/TWOCdemonstrator/tree/main/tools. Accessed: 2020-04-19

[23] Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative rules for linked data generation at your fingertips! In: Gangemi, A., Gentile, A.L., Nuzzolese, A.G., Rudolph, S., Maleshkova, M., Paulheim, H., Pan, J.Z., Alam, M. (eds.) The Semantic Web: ESWC 2018 Satellite Events, pp. 213–217. Springer, Cham (2018)

medRxiv preprint doi: https://doi.org/10.1101/2023.04.23.23289000; this version posted November 16, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[24] Phenopackets RDF Sschema. https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema. Accessed: 2020-04-19

[25] ISA tools environment. https://github.com/Xomics/Isatools_environment. Accessed: 2020-04-19

[26] ISA tools API. https://isa-tools.org/isa-api/content/index.html. Accessed: 2020-04-19

[27] The FAIR Data Point in CMBI. https://fdp.cmbi.umcn.nl. Accessed: 2020-04-19

[28] Digital Research Environment. https://www.radboudumc.nl/en/research/radboud-technology-centers/data-stewardship/digital-research-environment. Accessed: 2020-04-19

[29] Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L.A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J.D., Törnroos, J., Vasallo, C., Veal, C.D., Brookes, A.J.: Beacon v2 and beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Human Mutation **43**(6), 791–799 (2022) https://doi.org/10.1002/humu.24369 https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.24369