Multi-disciplinary Research: Open Science Data Lake

Vincent-Nam Dang, Nathalie Aussenac-Gilles, Franck Ravat

Multi-disciplinary research: Open Science Data Lake

Vincent-Nam Dang[1,2], Nathalie Aussenac-Gilles[1][0000−0003−3653−3223], and Franck Ravat[1,2][0000−0003−4820−841X]

[1] IRIT, CNRS, Université de Toulouse, France
[2] Université Toulouse Capitole, France

Abstract. Open Science aims to establish an interdisciplinary exchange between researchers through knowledge sharing and open data. However, this interdisciplinary exchange requires exchanges between different research domains and there is currently no simple computerized solution to this problem. Although the data lake adapts well to the constraints of variety and volume offered by the Open Science context, it is necessary to adapt this solution to (1) the accompaniment of data with metadata having a specific metadata model depending on the domain and community of origin, (2) the cohabitation of open and closed data within the same open data management platform, and (3) a wide diversity of pre-existing research data management platforms to deal with. We propose to define the Open Science Data Lake (OSDL) by adapting the Data Lake to this particular context and allowing interoperability with pre-existing research data management platforms. We propose a functional architecture that integrates multi-model metadata management, virtual integration of externally stored (meta)data and security mechanisms to manage the openness of the platforms and data. We propose an open-source and plug-and-play technical architecture that makes adoption as easy as possible. We set up a proof-of-concept experiment to evaluate our solution with different users from the research community and show that OSDL can meet the needs of transparent multidisciplinary data research.

Keywords: Interdisciplinarity · Open Science · Data Lake

1 Introduction

The need for interdisciplinarity in research is growing [1]. This need is expressed through the increasing efforts to implement Open Science (OSci). Data management solutions exist within research communities to handle community-specific data. However, interdisciplinarity brings in new challenges with the management of a wide variety of research data and a need for data openness. The establishment of bridges between communities creates a different context for the design of new solutions. New actors, with their own knowledge and needs, are emerging in relation to intra-community solutions. New contexts also emerge creating additional constraints to ensure that needs are always met. There is a need to

manage the cohabitation of open and closed data or the management of a wider variety of data and needs around this data, notably with metadata or processing. Specifically in the case of OSci, there are several additional challenges [15]: (1) the need for interoperability with a wide variety of existing data management solutions, (2) data and metadata format issues, (3) a rapid increase in the volume of data generated, both batch data and stream data, or even real-time data, (4) the need for significant time and resources for the implementation of common standards or metadata models. The data lake is a big data analytics solution that addresses the wide variety and volume of data. The data lake has become popular in research data management projects that mix several communities (EOSC with ESCAPE [7], Data Terra with Gaia data project[3], ESA / NASA with MAAP [4], European Commission with Destination Earth [8]). However, Open Big Data is a specific context that brings many additional constraints. We propose a new functional and technical data lake architecture adapted to the OSci context and evaluated by experimentation: the Open Science Data Lake.

In part 2, we explore the different OSci data management platforms and the place of datalakes within them. In part 3, we propose a functional architecture detailing the important additions to transform a multi-zone data lake architecture into an OSDL. In part 4, we propose a plug-and-play and open-source technical architecture. In part 5, we evaluate our solution through a proof of concept evaluated by users and compared to 3 existing data set search platforms.

2 Related works

Open Science is made up of a large number of data management platforms of all types. More than 3,000 platforms are listed on Re3data[4]. These platforms can be diverse, depending on the type and theme of the data, the volume or the community needs. These platforms can be based on noSQL databases, such as MongoDB [18], domain- or data-type-specific databases [16], data-warehouses[5], catalog-type web applications[6], specific solutions such as Dataverse[7], or many others. However, these solutions all have their limitations: a lack of scalability of interoperability with other platforms, a lack of variety in analyses or the type of data that can be managed, a lack of openness and others reasons.

The need to unify data access points to offer greater richness in data retrieval is growing. For this reason, more and more projects are based on data lakes[8]. This big data analysis solution meets a wide range of analysis and data volume management needs. It can be adapted to all fields and all types of data, whether in physics [3], medicine [11] or biology [13]. Data lake architectures have evolved over time [10, 14]. Initially intended as a raw data storage area, other functional areas have been integrated to meet more needs, including data processing and metadata management. However, these architectures are designed to manage models with a fixed metadata model, in which metadata will be generated during the data life cycle in the data lake. As it stands, managing pre-existing metadata

[3] www.gaia-data.org [4] www.re3data.org/ [5] www.biosino.org/bmdc/aboutUs/organization [6] re3data.org [7] https://ada.edu.au/ [8] data.openei.org/data_lakes

is not part of the data lake context. This is an obstacle to managing the variety present in OSci. In order to move forward with OSci, the FAIR Principles help define the directions in which this information sharing can take place [17]. With regard to the FAIR principles, the data lake lack of mechanism to meet the I3 principle, which concerns the interconnection of metadata. More focused on interoperability [5], the data lake does not functionally possess the mechanisms needed to be interoperable with other platforms. However, this is not a trivial issue. There are over 1600 standards[9] for metadata definition, including models, guidelines or terminology artifacts [12]. These different standards continue to evolve and expand with the adoption of OSci.

3 OSDL : Functional architecture

The number of asset profiles in OSci is enriched compared with the classic data lake context [9]. There is a whole gradient of data types, from internal data to open data. The opening up of data and platforms creates the presence of users external to the initial context of the platforms. Approaching the problem of OSci as a whole requires to take these assets into account, as well as the large volume and variety of data from OSci. But it is also necessary to integrate the wide variety of pre-existing system assets for data management. Designing

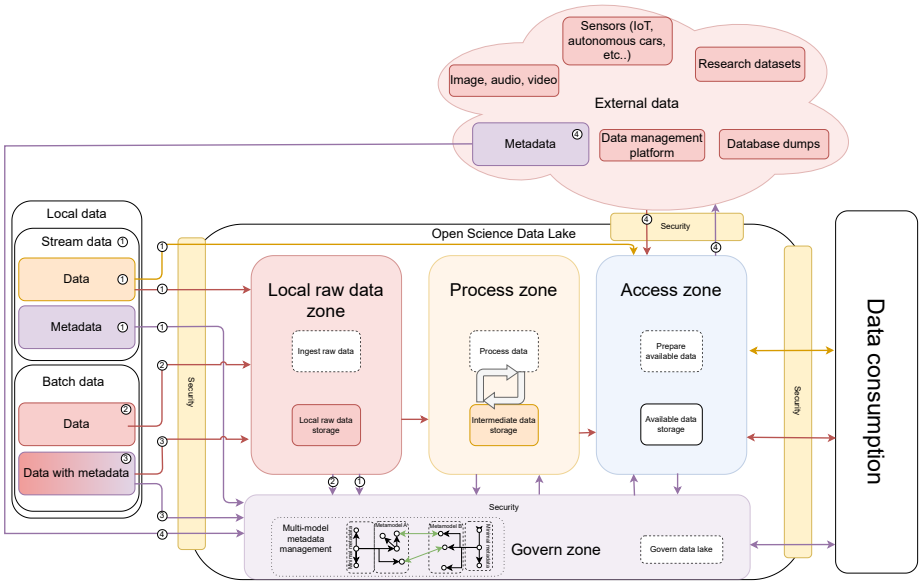

Fig. 1. Functional architecture

[9] fairsharing.org/

an Open Big Data [2] solution requires taking into account 2 major aspects, in addition to the constituents of a Big Data solution. Many data and data management solutions already exist. We need to integrate these pre-existing data and enable interoperability with pre-existing data management platforms in OSci. In addition, the enrichment of the assets to be managed, compared to a usual Big Data context, requires the design of security mechanisms as a core object of the architecture to protect against the associated threats specific to OSci [9]. With regard to the FAIR Principles, we need to address the issue of interoperability. We propose a functional architecture of OSDL (see Fig. 1) where we find the 4 main zones of a multi-zone data lake [6]: the raw data zone ingests the data in the original format, the process zone allows the implementation of treatments on the data, the access zone allows the access and consumption of the processed data and the governance zone contains the metadata as well as the governance mechanisms of the data lake. We observe a new type of storage to be integrated into the OSDL architecture: **external storage**, i.e. external data is stored in existing data management platforms. The volume of OSci data does not allow to copy, store and manage it as local data. This new type of storage requires the ability to manage data and metadata acquisition protocols from data management platforms. Metadata can be used to index large volumes of data. However, it is necessary to integrate the possibility of retrieving metadata only when it is needed, to avoid an explosion in metadata volume. In addition to the two usual profiles (batch and stream data), external storage creates two new data profiles with batch data accompanied by metadata and metadata alone to be ingested. Fig. 1 illustrates stream data with orange arrows, batch data with red arrows and metadata with purple arrows.

- Data profile 1 consists of stream data, possibly with temporal constraints. Once the stream has been initialized and the corresponding metadata ingested, the data directly arrives in the access zone, where it is consumed in the shortest possible time.
- Data profile 2 consists of batch data. This data is received and inserted into the raw data area. Metadata is generated as the data passes through the various OSDL zones [10], allowing the life cycle of the data to be monitored.
- Data profile 3 is made up of batch data accompanied by predefined metadata with a specific model. The data is inserted in raw data zone. In parallel, metadata are inserted without modifications in the governance zone.
- Data profile 4 consists of data stored externally to the OSDL platform. Only the metadata is ingested into the OSDL to allow the knowledge of the associated data. Data can be queried and used in a similar way to other data profiles, without being stored locally.

3.1 OSDL: Interoperability

To support external data storage, exchanges with other data management platforms have to be handled. This requires interoperability between platforms and OSDL. We take as our definition of interoperability the one we proposed in a

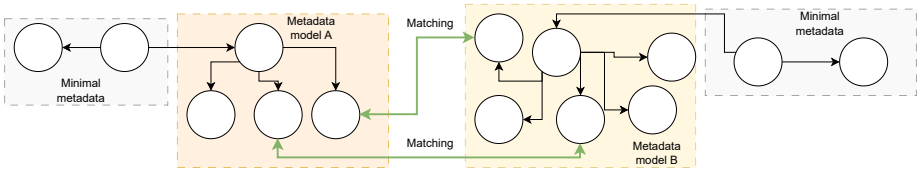

Fig. 2. OSDL metadata management: multi-model with matching

previous article [5]. We aim to enable the exchange of usable information on the different datasets. The data to be exchanged is the dataset metadata, and the useful information is the one about this metadata, the so-called metadata models. For the sake of simplicity, we deal with the 2 layer categories: system layers and process layers. For system layers, we chose to use a REST API to enable communications. In Re3Data.org, the REST API is the most widespread type of API among data management platforms, with almost 45% of platforms having communicated information about their API to Re3Data (interoperability by standardization). In addition to standardization with a large number of platforms, REST API technologies enable simple interfacing with a wide range of existing communication technologies (interoperability by gateway implementation). For process layers, we proposed to adopt multi-model metadata management (Fig. 2). This requires to handle matchings between models. Multi-model management means that external metadata can be stored, but also that these metadata can be used to query external platforms. In this way, metadata can be retrieved when needed, rather than stored locally; and no pivot model is required. We have explored interoperability and matchings more in depth in a former paper [5].

3.2 OSDL: Data security

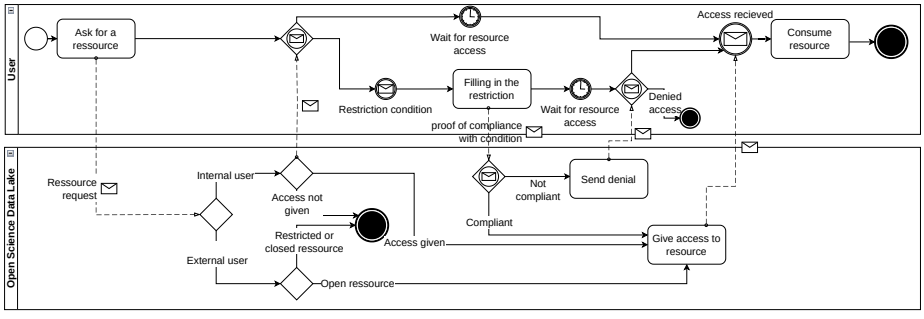

Fig. 3. Access control on OSDL resource process

To avoid any loss of data, trust or time for researchers, security mechanisms are necessary [9]. Access control to OSDL resources is integrated into all the platform pipelines (see Fig. 3). These access controls, combined with user, group and project management, make it possible to set up privileges for different resources (see Fig. 4). This allows different asset categories to be set up, and assets to be logically secured as required. These mechanisms ensure legal compliance with licenses, based on Principle R1.1 of the FAIR Principles.

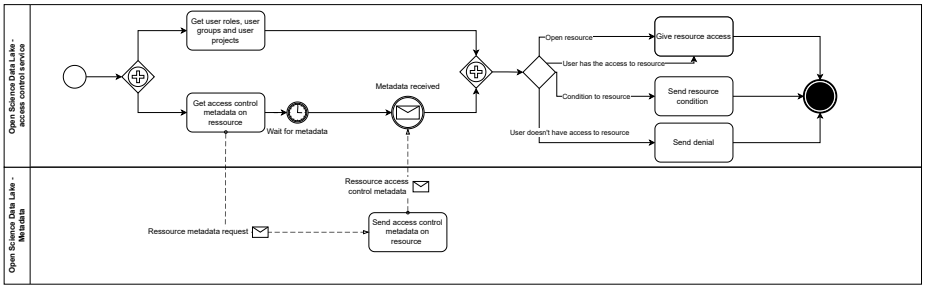

Fig. 4. Access to user privileges and required privileges for a resource

4 OSDL: Technical architecture

OSDL must also be technically adapted to OSci. For this aspect, this architecture must be an Open Source solution [2]. However, to avoid the durability issues encountered in Dataverse[10], this architecture needs to be modular and easy to be maintained by external developers. In addition, mechanisms must be devised to ensure an adoption as wide and simple as possible. We propose an open-source implementation (Fig. 5) of the OSDL (the code is available in a git Repository [11]). We chose tools by considering the longevity, the openness of code and the use of REST APIs for interacting with them in a concern of simplicity, use, maintenance and interoperability. The entire architecture has been designed as containerized, using Docker containers. Automatic deployment tools have been developed to allow a one command deployment on most servers.

This technical solution is an adaptation of the architecture proposed in a previous paper [6] to the context of OSci. Data processes are managed by Apache Airflow. This tool enables workflows to be managed in the form of Directed Acyclic Diagrams. This makes it easy to track all operations in a processing chain. Other tools can be called up in the process data area by Apache Airflow for more specific processes. The management of raw data, transformed data and data processing pipelines is more detailed in this paper [6]. For security management, added security mechanisms are integrated into the REST API, providing

[10] dataverse.org/presentations/open-monolith-keeping-your-codebase-and-your-headaches-small [11] github.com/vincentnam/docker_datalake

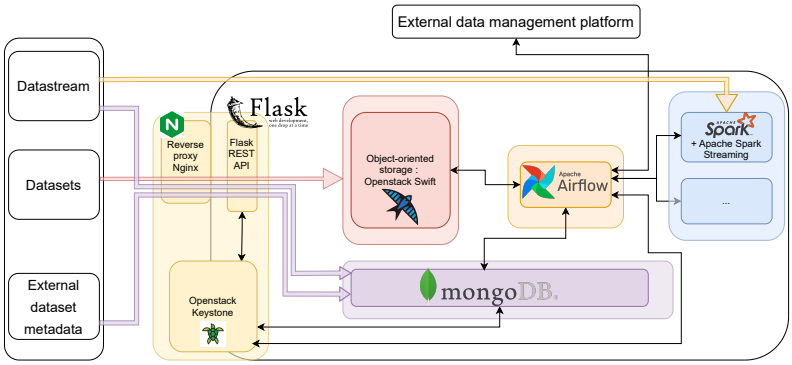

Fig. 5. OSDL: Technical architecture

a single access point to all OSDL resources. User management is implemented with Openstack Keystone, based on legal information in metadata.

Multi-model management is implemented in a MongoDB database. Models and matchings between models are stored independently in different collections. MongoDB allows to take advantage of noSQL flexibility on models and to handle JSON-LD for semantic metadata and linked data without functional redundancy with other data lake services. Moreover, the native interoperability by standardization with REST APIs eliminates pre-processing operations on received messages and reduces the load. The document format allows us to keep the list of matched keys for each model, so that match requests can be simple selections. From a technical point of view, the metadata management tool must be able to store and query metadata. Other needs are met by other data lake services (such as quality assurance pipelines with Airflow). Based on this, MongoDB is not a composite service (like OpenMetadata[12] or Opendatadiscovery[13], which relies on external database services and ElasticSearch) or based on a particular technology (like Apache Atlas[14], which relies on Hadoop). Since the solution meets our needs, this simplicity ensures lower maintenance and development costs. These aspects are essential to ensure that the solution is sustainable and that the problems encountered with monolithic solutions are not transposed to modular solutions. This is a major aspect of the solution's adoption in OSci.

5 Evaluation

We have set up an experimental implementation of a proof of concept of OSDL[15] (see Git repository for an in-depth technical view). The aim is to evaluate the time saved by the user, the ability of OSDL to adapt to user needs, and the ability to implement a unified tool for cross-community access to research data with OSDL. We have selected metadata from 3 platforms from different domains

[12] https://open-metadata.org/ [13] opendatadiscovery.org [14] https://atlas.apache.org/ [15] anonymous.4open.science/r/opendatalake_expe-6522

	R1	R2	R3	R4	R5	R6	R7	R8
AERIS (A)			X	X			X	
ODATIS (O)				X*			X	
RCSB PDB (P)	X	X			X	**		
OSDL (POC)	X	X	X	X	X	X	X	X

Table 1. Request availability by platform ; X: Request can be made in this platform.

*: Queries on the geolocation of ODATIS data are made through a map. This type of query tool does not allow precise queries; **: PDB allows searches on entire journal titles, but it is not possible to make sub-string research in titles.

and communities with different metadata models (AERIS[16], ODATIS[17], RCSB PDB[18]). Cross-platform communication mechanisms were simulated by integrating metadata into a single database, due to the lack of a method for scripting communications with all platforms. Matchings between models were integrated into a specific collection, and POC queries were sent to our database on a metadata path as well as on all equivalent metadata in the matching. We designed a set of 8 queries on metadata to specify search across multiple attributes, including natural phenomena and protein data (described in Github repository). We set up an experimental scenario with 11 users that were asked to execute the 8 queries on the 4 data retrieval platforms (AERIS, ODATIS, RCSB PDB and the OSDL proof of concept). We selected users so as to approximate the distribution proposed in a study on OSci (cf. Q12[19]) with 3 categories of comfort with open dataset search platforms that we assimilate as equivalent to those in the study: comfortable (≈20%), somewhat comfortable (≈40%) and not comfortable (≈40%). The users have not been trained to use the platforms (in order AERIS, ODATIS, RCSB PDB and finally the POC of OSDL). We measured the time required by each user to perform queries on each platform.

Mean time for request (in second)	AERIS	ODATIS	RCSB PDB	OSDL
Without error	26.74	22.73	31.08	22.96
With error	27.84	21.67	34.32	22.93

Table 2. Request mean time for each platform

We have observed that OSDL enables a greater variety of metadata requests (see Table 1) thanks to the richness provided by multi-model management associated with matchings between these models. To manage the models of the 3 platforms, we had to set up two JSON documents weighing a total of 3.3Kb. This theoretically allows us to retrieve information from almost 200 different platforms present on Re3data having implemented ISO 19115 (the model implemented on the ODATIS platform).

[16] www.aeris-data.fr [17] www.odatis-ocean.fr [18] www.rcsb.org [19] map.scnat.ch/en/activities/open_data_survey

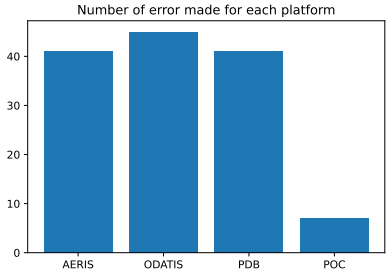

Fig. 6. Number of errors in user querying, for a total of 88 queries per platform

We have found that OSDL provides a data retrieval tool with average usage times at least equivalent to other platforms (see Table 2), while at the same time providing tools that are simpler to use and more user-friendly (see Fig.6). We managed to integrate data from existing OSci platforms without the need to modify existing platforms. OSDL is interoperable with other pre-existing platforms without other specific requirements than a (meta)data acquisition mechanism.

6 Conclusion

The specificities brought to Big Data by Open Science (OSci) mean that new constraints must be taken into account, with the arrival of new assets. Interoperability and data security are 2 new components to be integrated into the very heart of Open Big Data solution design. We have proposed a data lake architecture adapted to OSci: the Open Science Data Lake (OSDL). Its novel architecture is based on recognized data lake architectures, enabling (i) local data integration by adding (ii) external data storage management for interoperation with existing OSci data management solutions, and (iii) security mechanisms at the very heart of the architecture to guard as far as possible against loss of data, trust or time in the research knowledge creation process. We carried out a POC which we evaluated through an experiment with users from the world of scientific research. This evaluation enabled us to show that OSDL saves time and broadens the scope of data retrieval by researchers. By design, OSDL's allows integration of metadata from other platforms without any additional workloads for the other platforms. With regard to the FAIR principles, our solution meets principles 1 and 3 of metadata Interoperability, which is a necessary but not sufficient step towards data interoperability, and all the layers of interoperability [5]. Further work will focus on adding mechanisms to enable scaling-up through automation of meta-metadata exchanges, by designing of a federation of OSci data management platforms.

References

1. Barry, A., et al.: Logics of interdisciplinarity. Economy and society **37**(1) (2008)
2. Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P., Görögh, E., Helbig, K., Kramer, B., Labastida, I., Niemeyer, K., Psomopoulos, F., Ross-Hellauer, T., Schneider, R., Tennant, J., Verbakel, E., Brinken, H., Heller, L.: Open Science Training Handbook. Zenodo (Apr 2018). https://doi.org/10.5281/zenodo.1212496
3. Bird, I., et al.: Architecture and prototype of a wlcg data lake for hl-lhc. In: EPJ Web of Conferences. vol. 214, p. 04024. EDP Sciences (2019)
4. Bugbee, K., et al.: Advancing open science through innovative data system solutions: The joint esa-nasa multi-mission algorithm and analysis platform (maap)'s data ecosystem. In: IGARSS 2020 - IEEE International Geoscience and Remote Sensing Symposium. pp. 3097–3100. IEEE (2020)
5. Dang, V.N., Aussenac-Gilles, N., Megdiche, I., Ravat, F.: Interoperability of open science metadata: What about the reality? In: International Conference on Research Challenges in Information Science. pp. 467–482. Springer (2023)
6. Dang, V.N., Zhao, Y., Megdiche, I., Ravat, F.: A zone-based data lake architecture for iot, small and big data. In: 25th International Database Engineering & Applications Symposium (IDEAS 2021) (2021)
7. Di Maria, R., Dona, R.: Escape data lake. In: EPJ Web of Conferences. vol. 251. EDP Sciences (2021)
8. Juarez, J.D., Schick, M., Puechmaille, D., Stoicescu, M., Saulyak, B.: Destination earth data lake. Tech. rep., Copernicus Meetings (2023)
9. Peisert, S., Welch, V., Adams, A., Bevier, R., Dopheide, M., LeDuc, R., Meunier, P., Schwab, S., Stocks, K.: Open science cyber risk profile (oscrp), version 1.3.3 (2017). https://doi.org/DOI: 10.5281/zenodo.7268749
10. Ravat, F., Zhao, Y.: Data lakes: Trends and perspectives. In: Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30. pp. 304–313. Springer (2019)
11. Ren, P., et al.: Mhdp: an efficient data lake platform for medical multi-source heterogeneous data. In: Web Information Systems and Applications: 18th International Conference, WISA 2021, Kaifeng, China. pp. 727–738. Springer (2021)
12. Sansone, S.A., et al.: Fairsharing as a community approach to standards, repositories and policies. Nature biotechnology **37**(4), 358–367 (2019)
13. Sarramia, D., Claude, A., Ogereau, F., Mezhoud, J., Mailhot, G.: Ceba: A data lake for data sharing and environmental monitoring. Sensors **22**(7), 2733 (2022)
14. Sawadogo, P., Darmont, J.: On data lake architectures and metadata management. Journal of Intelligent Information Systems **56**, 97–120 (2021)
15. Tanhua, T., et al.: Ocean fair data services. Frontiers in Marine Science **6**, 440 (2019)
16. Wang, Y., Ling, Y., Gong, J., Zhao, X., Zhou, H., Xie, B., Lou, H., Zhuang, X., Jin, L., Initiative, H., Fan, S., Zhang, G., Xu, S.: Pgg. sv: a whole-genome-sequencing-based structural variant resource and data analysis platform. Nucleic Acids Research **51**(D1), D1109–D1116 (2023)
17. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
18. Zhou, C., Xu, Q., He, S., Ye, W., Cao, R., Wang, P., Ling, Y., Yan, X., Wang, Q., Zhang, G.: Gtdb: an integrated resource for glycosyltransferase sequences and annotations. Database **2020** (2020)