A data mesh for research data management

Claudio Cacciari
SURF
Moreelsepark 48,
3511 EP Utrecht,
Netherlands
claudio.cacciari@surf.nl

ABSTRACT

In our experience, as data management team of SURF (the Dutch national IT cooperative for education and research), we have seen iRODS adopted successfully in various organizations, like universities or medical centers, or part of them, like departments or laboratories. It is typically used as the "brain" of a data infrastructure to implement storage virtualization, storage tiering and, in general, the full data lifecycle management. However, when we look at research projects that encompass multiple organizations, or multiple units within big, distributed organizations, we notice that there are more difficulties, both of human and technical nature. Sometimes just opening ports in a firewall is quite hard, or, in other cases, the data governance is not well defined or understood. In the last two decades, in that part of the data management field more business oriented, new paradigms and concepts emerged, like data warehouse, data lake and, more recently, data fabric, data product and data mesh. In this presentation we borrow some of those concepts and we map them to research data infrastructures and iRODS to propose a solution that addresses some of the issues of building data platforms for large, distributed infrastructures. We will describe an initial technical and organizational implementation with a perspective for the next steps.

Keywords

Data Mesh, Data Product, FAIR Data, Neptune API, Research Data Management

INTRODUCTION

In our experience, as data management team of SURF (the Dutch national IT cooperative for education and research), we have seen iRODS adopted successfully in various organizations, like universities or medical centers, or part of them, like departments or laboratories. It is typically used as the "brain" of a data infrastructure to implement storage virtualization, storage tiering and, in general, the full data lifecycle management. However, when we look at research projects that encompass multiple organizations, or multiple units within big, distributed organizations, we notice that there are more difficulties, both of human and technical nature. Sometimes just opening ports in a firewall is quite hard, or, in other cases, the data governance is not well defined or understood. How can we make the creation of project or community-oriented data infrastructures easier? How can we make them benefit from existing iRODS and YODA services already well established within partner organizations?

RESEARCH DATA MESH

The reproducibility of research data was and still is a problem, named the reproducibility crisis. The concept of FAIR data was created to address it, providing guidelines to improve Findability, Accessibility, Interoperability, and Reuse of digital assets [1], or, in other words, to improve the quality of the research data.

Findability, accessibility and interoperability of research data have been improved over the past years, through various technologies; however, there are still important gaps in the data infrastructures. Often, they require a level of technical expertise which is not common among the researchers and there is a lack of interoperability among the tools. The result is that the users are forced to create data workflows oriented by the technology or the service which they use, rather than the domain which they belong to.

In the field of business data, a similar problem was described a few years ago [2]. The existing data architectures, Data Warehouse and Data Lake, were considered in a state of crisis, because they were not up to the growing complexity of the data pipelines and of the organizational needs. Therefore, a paradigm shift was proposed to take up the new challenges. It was called Data Mesh and its main purpose is to create a decentralized data architecture that enables the extraction of large-scale analytical data [3].

We propose here to adapt the Data Mesh concept for the field of research data, where its scope can be re-defined in this way: to create a decentralized data architecture that enables the exchange of large-scale FAIR data.

There are four underpinning principles that any data mesh implementation embodies to achieve the promise of scale, while delivering quality and integrity guarantees needed to make data usable:

1. *domain-oriented decentralized data ownership and architecture,*
2. *data as a product,*
3. *self-serve data infrastructure as a platform, and*
4. *federated computational governance.*

Where "federated computational governance" means *a governance model that embraces decentralization and domain self-sovereignty, interoperability through global standardization, a dynamic topology and most importantly automated execution of decisions by the platform* [2].

The first point helps us to address the problem of the fragmentation of data infrastructures among different technologies and services. The idea is to support a different way to provide and consume the data, aggregating them per research domain, community, project.

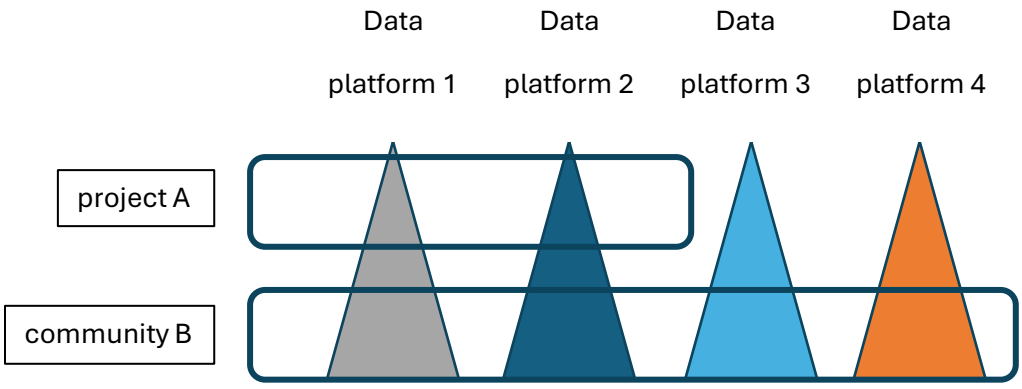

Figure 1. Providing data per domain or project.

In this way, a researcher of project A can focus on working on the project's data, not on data platform 1 and 2. The federated approach breaks silos, which are built around an organization, or a part of it, or around functions of the data life cycle. For example, certain data services provide sharing capabilities, but they are not suitable for long term preservation or viceversa.

A data product is defined by some principles and, as someone noticed [4][5], they are closely related to FAIR data principles.

FAIR	Data mesh characteristics	Why it is important
Findable	Discoverable, addressable	Data product teams can find the data products from other teams quickly and independently
Accessible	Addressable, self-describing, secure, natively accessible	Data product teams can consume the data products from other teams without significant friction or a lengthy process
Interoperable	Trustworthy, secure, self-describing, interoperable, addressable	Data products can be consistently and reliably combined together, they follow standardization and harmonization rules, leading to greater usage
Reusable	Trustworthy, secure, self-describing, interoperable	Data products can serve more than one use case, and are re-used in a new setting creating more value
	Valuable on its own	A data product should carry a dataset that is valuable and meaningful on its own-without being joined and correlated with other data products

Figure 2. FAIR data as Data Products [5].

In the business field, the minimal size of a unit of a data mesh is the data product, so a node in the mesh is a data product. In the research field, we may have similar scenarios, but we have certainly also cases where the Data Mesh node is a set of data products, aggregated by project or scientific domain. An interesting example of such a node is that of a data common, *a cloud-based data platform with a governance structure that allows a community to manage, analyze and share its data* [6].

The third principle, being a self-serve data infrastructure as a platform, gives us two hints. The word "self-serve" indicates the importance of the data consumers. A Data Mesh democratizes the data giving more power to the researchers in discovering and accessing the data without the intermediation of a central team. While the sentence "data infrastructure as a platform" defines the Data Mesh itself as a platform, thus opening the possibility to create a Data Mesh of Data Meshes, where a Data Mesh is a node of a Data Mesh of higher level.

Finally, the last point reminds us that, even if it is a decentralized system, it must have a common governance. The federated approach allows each data owner and provider to maintain control over their own data, dictating the conditions to get access. These conditions have to be harmonized (i.e. not conflicting) anyway with the common Data Mesh rules and automated to support the self-serve capability.

The possibility to use a Data Mesh as a building block for distributed systems has been described also in relation to a Data Space. Data Spaces are *a distributed and standards-based approach to enabling data sharing and use across organizations* [7]. Data governance, authorization and connection capabilities of Data Spaces complement the Data Mesh capabilities, which can be seen as a Data Space in miniature.

IRODS AS A RESEARCH DATA MESH NODE

Given the description of a Research Data Mesh offered in the previous chapter, iRODS and YODA can be seen as nodes of a Research Data Mesh. Then we could offer a solution to the initial problem of the creation of project or community-oriented data infrastructures, building a mesh where the services and the data of the different partner organizations are shared in a FAIR way. The federated governance approach would require an organizational effort to define common rules, but minimal, just enough to define the "borders" of the mesh's space, leaving each partner in control of its own data.

Hub-and-Spoke architecture

Sometime, an organization has multiple services that are involved in the data lifecycle, not just a single iRODS or YODA instance. We could apply the Research Data Mesh paradigm within the organization too. Once connected to the external Research Data Mesh, we would realize a mesh of meshes. However, the more the mesh is stratified in multiple layers, the harder is the challenge of the federated governance. We propose to mitigate that complexity, adopting a hybrid mesh architecture. Where the external Research Data Mesh is coupled with a hub-and-spoke architecture within the organization. A hub-and-spoke design provides a central point, the hub, where it is possible to apply the governance rules and, in general, organization-wide data policies. This central point would represent the whole organization as a node in the external Research Data Mesh, guaranteeing consistency in the exchange of data and metadata between the organization and the other nodes of the mesh.

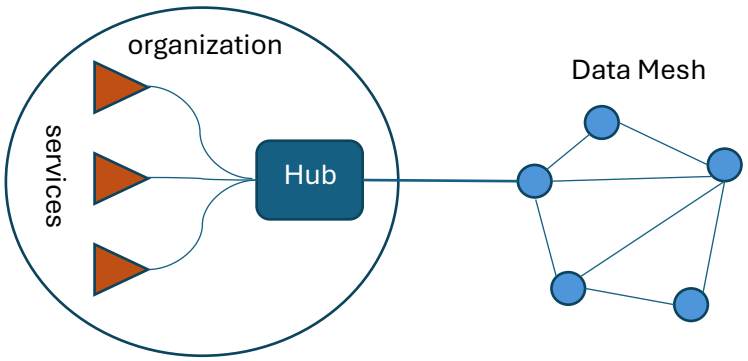

Figure 3. Hybrid data mesh.

However, the hub in the hub-and-spoke architecture could become a bottleneck and a single point of failure. A redundant implementation can mitigate the latter weakness, but in order to remove the former one, we propose limiting the scope of the hub to metadata. In fact, any data workflow can be designed so that it is metadata driven, hence if we keep the metadata consistent, the governance and data policies rules will be consistent as well. On the other hand, the data will be transferred through a direct connection to or from the service hosting them, so that the performance and the scalability will be optimized.

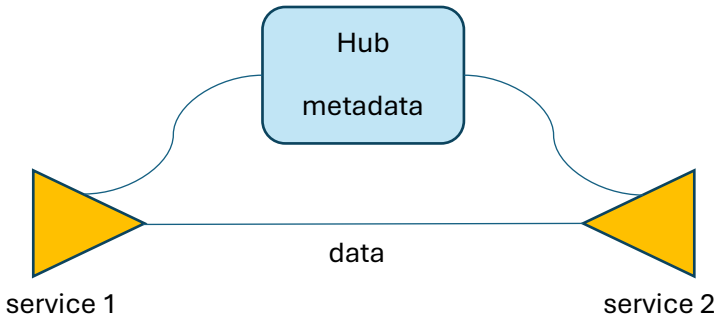

Figure 4. Hub for metadata only.

Research Data Hybrid Mesh node

Which features should offer the node of this Research Data Hybrid Mesh, being at the same time the hub of a hub-and-spoke topology?

We propose that it is defined by the following properties:

1. It is a service registry: it needs to be aware of all the services within the organization
    a. to advertise them to the other nodes of the mesh.
    b. To support the creation of data workflows that use different services.
2. It is a user registry: it needs to identify any "actor" that interacts with it to support decisions based on identity and role.
3. It is a metadata catalog: the metadata related to users, providers and data are necessary to enforce policies based on them, automate as much as possible the decisions and to fulfill the definition of data product.
4. It is an API gateway: it needs to route and control the API requests between the organization domain and the external world.
5. It is a data sharing point: where the available data can be advertised to other users, providers, nodes ... And where data can be discovered through metadata and requested from the data provider hosting them. As explained above, the data should not be transferred through the node, but directly from/to the data provider.

These properties are familiar to those who have some knowledge of iRODS, because they overlap largely with its features. Therefore, in the case that an instance of iRODS, or YODA is the only service connected to the Research Data Mesh, the hub-and-spoke architecture could "collapse" on iRODS itself, while in case of multiple services an additional logical layer, the hub, is needed.

CONCLUSION

We have identified fragmentation, from a technological and organizational point of view, as one of the main obstacles hindering the creation and effectiveness of large data infrastructures. We have adapted a paradigm and an architecture, the Data Mesh, born in the business field, to the research data field and we have shown that it can help in overcome the fragmentation issue, democratizing and simplifying the access to the data. iRODS (and YODA) can be one of the building blocks of such Research Data Mesh. Generalizing the solution for organizations with multiple data services, we can design the node of the mesh as a hub of and-and-spoke architecture, so that multiple data services are connected to the mesh via the hub. In this case, iRODS would be one of the services connected to the hub. To support the proposed solution, the hub should act as:

1. A service registry.
2. A user registry.
3. A metadata catalog.

4. An API gateway.
5. A data sharing point.

REFERENCES

(1) https://www.go-fair.org/fair-principles/ (visited on May 9th 2025)
(2) https://martinfowler.com/articles/data-mesh-principles.html (visited on May 9th 2025)
(3) Inês Araújo Machado, Carlos Costa, Maribel Yasmina Santos,
Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures,
https://doi.org/10.1016/j.procs.2021.12.013.
(https://www.sciencedirect.com/science/article/pii/S1877050921022365)
(4) https://www.linkedin.com/pulse/fair-guiding-principles-data-mesh-jeroen-angenent/ (visited on May 9th 2025)
(5) https://www.thoughtworks.com/insights/blog/data-strategy/data-mesh--helping-life-sciences-organizations-get-more-from-the (visited on May 9th 2025)
(6) Grossman R. L. (2023). Ten lessons for data sharing with a data commons. Scientific data, 10(1), 120. https://doi.org/10.1038/s41597-023-02029-x
(7) Antti Poikola (Sitra), P J Laszkowicz, Ville Takanen and Teemu Toivonen (Futurice) (2023), Technology Landscape of Data Spaces. Sitra Publisher.
(https://www.sitra.fi/en/publications/technology-landscape-of-data-spaces)