Enabling interdisciplinary research in Open Science: Open Science Data Network

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

Enabling interdisciplinary research in Open Science: Open Science Data Network

Vincent-Nam Dang[1], Nathalie Aussenac-Gilles[2 [0000−0003−3653−3223]], Imen Megdiche[3 [0000−0002−1331−8662]], and Franck Ravat[1 [0000−0003−4820−841X]]

[1] IRIT, CNRS (UMR 5505), Université Toulouse Capitole, France
[2] IRIT, CNRS (UMR 5505), France
[3] IRIT, CNRS (UMR 5505), INU Champollion, ISIS Castres, Université de Toulouse

Abstract. The aim of Open Science is to open up data to enrich knowledge creation processes. At present, Open Science actors face problems when trying to find and exchange data. Research data management platforms need to address the issue of interoperability to enable interdisciplinary research. Some solutions are available for specific communities, but none addresses the problem as a whole. Based on an extension of the theoretical model of interoperability, which enables us to define the criteria for an information exchange, we have quantitatively evaluated information exchange in Open Science. On the basis of this explorative study, we propose an inter-community and inter-disciplinary information exchange network solution enabling decentralised and federated data management as well as a unified search for datasets across all the entities registered in this network: the Open Science Data Network (OSDN). We carried out a proof of concept to assess the feasibility of the solution. We also evaluated this solution by applying it to a completed agronomy research project. This evaluation enabled us to measure a 7% increase in the volume of data, with an 80% reduction in the time needed to find this data. In addition, users have been able to design new intra- and interdisciplinary futures works with data found.

Keywords: Information System · Interoperability · Data Integration · Metadata Management · Open Science

1 Introduction

Open Science is a global research movement aimed at opening up knowledge creation processes to enable collaboration and enrichment of the creative process. Open Science actors describe a need for a inter and intra-community coordination [8] to accelerate the adoption of Open Science movement. The FAIR principles define the direction that should be taken to improve information and data sharing, particularly in the context of Open Science [34]. Findability in the sense of the FAIR principles is defined as the ease of finding data for both humans and machines [12]. Researchers from different fields and communities agree to say that there is a problem of findability of datasets, whether raw data or

datasets resulting from research work [15,11,22,25,20,4,27,5,19]. Open Science actors explain this problematic with 2 main reasons, mainly related with the dataset metadata models:

- the variety of the metadata models for these datasets alters findability. We find a large number of differents models, standardized [34] or specific to a platform [22]. Scientists have noticed a lack of interoperability between these models [25,20], therefore increasing the adoption cost of open data solutions.
- the lack of coordination between efforts [11] leads to a large number of research data management platforms, some of them being redundant [11,22,13,5], increasing the variety of used models. Users lack the resources – in time, manpower, knowledge and training – to get to grip with these models [25,20,19].

One way of addressing these issues is through a centralized data management platform in Open Science enabling harmonization of metadata models [30]. But several problems arise with centralisation: (i) impossibility to meet the specific needs of researchers from each different community [22,4]; (ii) too high volume for a single platform [30]; (iii) too costly to deploy such a platform that become a single point of failure [22]; (iv) no security, confidentiality or data control guarantee, whereas it is needed by researchers [25,20,13,19]. Centralisation is not viable to answer the Open Science information exchange problem. Several papers report that researchers claim the need for decentralized and federated data management platforms [22,30] with a unified access to metadata [22]. Several articles emphasize that to achieve an effective Open Science, the effort must be joint and community-based [25,20]. To put it another way, an appropriate solution should present the following features: (i) to manage the wide variety of metadata models from all Open Science communities and domains; (ii) to manage and coordinate with a decentralized solution the wide variety of pre-existing data management platforms, and take advantage of previous efforts; (iii) to provide an easy-to-adopt response to address users' lack of resources.

An intra- and inter-community coordination aims to exchange information to enable cooperation, which corresponds to interoperability [6,7]. However, before proposing a solution, it is needed to assess the type of required solution (infrastructure, user interface, communities, incentives or policy) [16] by getting a quantitative insight into the state of information exchange in Open Science. To the best of our knowledge, there is no quantitative evaluation available on the exchange of information.

We propose an extension of the theoretical model of interoperability [7]. This extension enables to explain the criteria for an information exchange in Open Science allowing us to propose a quantitative metric to assess the lack of information exchange. We then propose an architectural solution to address this lack of implementation of information exchange in Open Science: the Open Science Data Network (OSDN).

In section 2, we explore the notion of interoperability and Open Science data management solutions. In section 3, we extend the theoretical model of interoperability to apply this model to the quantitative evaluation of information exchange in Open Science. In section 4, we describe how the interoperability

of data management platforms is implemented in the OSDN and the network structuring of the OSDN. In section 5, we developed a Proof of Concept of OSDN. We made an user experiment with a researcher in agronomy and evaluated the contribution of our solution in terms of time savings, data set enrichment and new future work for this researcher.

2 Related works

Interoperability is a subject that regularly crops up in the scientific literature. A number of studies focused on various features of interoperability, leading to the distinction of several types of interoperability (technical, syntactical, semantic, platform, system, structural, conceptual, dynamic,..) [35,29,17,24]. In state of art on interoperability, we find two main components of interoperability: (i) the technical interoperability, which comes from the field of software engineering domain [33], the networks and telecommunication domain [32] or the database domain [14] with the technical problem of exchanging information between several information systems, (ii) the semantic interoperability [10,35] bringing in the need to add value to the information exchanged. The link between the OSI model and interoperability is made [23] when describing interoperability as a layered characteristic. Interoperability also take an important role in the Semantic Web. The Semantic Web contributes to establish interoperability between systems and people on the Web [28]. This objective is shared in Open Science, which focuses on research community [31]. A major focus was placed on semantic interoperability with the issue of knowledge and data exchange [35], in particular when modelling machine-processable metadata with standard vocabularies. This exploration of the different approaches to interoperability reveals a wide variety of approaches to interoperability, which makes it difficult to understand in a comprehensive and generalizable way.

In Open Science, there are many data management platforms. But there are few solutions for interoperating these data management platforms. OAI-PMH[4] is an information exchange protocol based on a harvesting mechanism. The harvester retrieves the information proposed by a service provider. This protocol has a star-shaped architecture. But this solution has its limitations when it comes to setting up inter-community exchanges across the whole of Open Science, linked to the use of a pivotal metadata model and the star-shaped structure of the network created by the protocol. There are other intra-community solutions for data management. Open Science Framework [21] is a general-purpose data management solution. However, the centralisation of the solution does not allow us to respond to all the problems of Open Science. The Beacon Network[5] has a larger scope in its technical conception. However, this solution is designed for a specific field: genetic mutations.

At present, we did not find a global solution for interdisciplinary and inter-community information exchange. To understand the challenges of such a solution, we proposed an unified theoritical framework [7] within which the many

[4] https://www.openarchives.org/pmh/ [5] https://beacon-network.org/

definitions of interoperability can be explained. In the following section, we complete this interoperability definition with additional characteristics of interoperability needed for an explanation of information exchange.

3 Open Science and Information exchange

Interdisciplinary information exchange is one of the acknowledged challenge in Open Science. It requires metadata interoperability. In this section, we explore the link between interoperability and information exchange (see Figure 1 as a guideline).

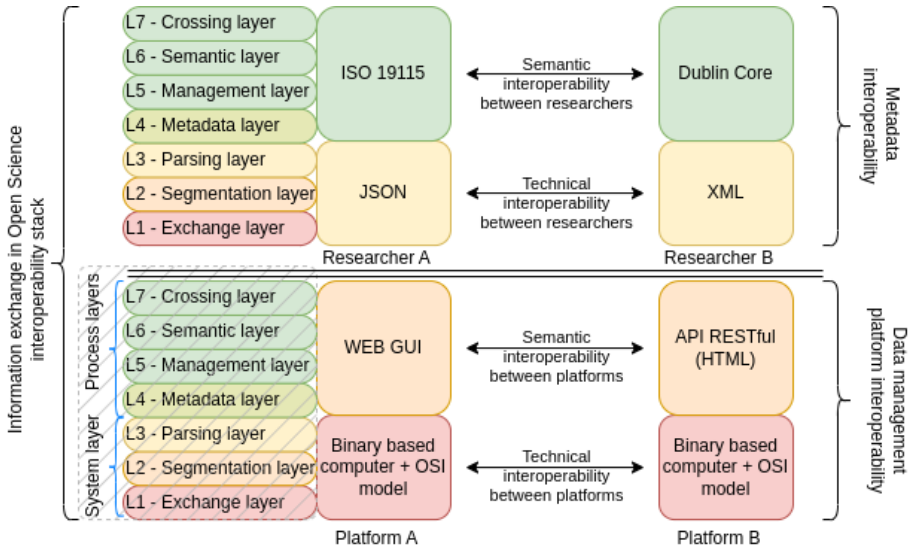


Fig. 1: Interoperability theoretical model applied to information exchange between researchers in Open Science

3.1 Formal notation

We have chosen to formalise the concept of data using formal grammar. We use "information" and "data" interchangeably because data do not differ structurally from the information [1]. We use the definition of a model as the representation of a domain conceptualization [9]. In the following, we assume that models are faithful representations of domain concepts and properties, allowing us to associate semantic to models. We use "platform" to describe any data management solution that can be used in Open Science to manage research data (like data catalog, data repository, databases, etc..).

In the rest of the paper, we use the notions of graph, formal grammar, probability and set theory, with the following notations:

- $g^{\text{L}}_{e_i}$, a formal grammar of information managed by entity e_i. This grammar defines information management rules in e_i associated to interoperability layer L, depending on the layer purpose.
- $l^{\text{L}}_{e_1}$, a formal language generated by $g^{\text{L}}_{e_1}$ as the set of words that respects rules of layer L.
- $G = (V, E)$, a graph where V is the set of nodes in the graph and E is the set of edges in the graph.
- $\mathcal{P}(\mathcal{S})$, the powerset of a set S, that we can describe as the set of all subsets of S.
- $Pr(X)$ the probability of an event X
- $|S|$, the cardinality of a set S
- $\mathbb{N}$, the set of natural numbers
- API and MD, respectively the set of API implemented by a platform and the metadata model implemented by a platform

3.2 Interoperability and information exchange

To understand the mechanisms and requirements for setting up interoperability, it is necessary to have a complete understanding of interoperability. We defined interoperability as *the ability of two communicating entities to work cooperatively through an exchange of information to achieve an objective* [7]. We define interoperability as a stack of seven layers (see hatched area in Figure 1) [7]. This stack can be divided into two groups of layers: system layers associated to **technical interoperability** and process layers associated to **semantic interoperability** (see blue braces in Figure 1) [7]. When the entire interoperability stack is implemented, **global interoperability** is then achieved.

Each layer covers a group of mechanisms to be implemented, which vary according to the **context**, the **objective** and the involved **communicating entities** [7]. As an example, the mechanisms of the parsing layer (layer 3) to be implemented for interoperability between two humans will focus on finding the words in a sentence. In the context of interoperability between two computer servers, it is necessary to distinguish the header from the payload.

We distinguish two types of interoperability mechanisms [7]. The first type covers **standardisation** [7], with the aim to establish a common and formal vocabulary for communicating entities in digital format. The second one covers mechanisms of **gateways implementation** [7]. The objective is to set up an entity or a mechanism to act as a bridge between tools of a specific interoperability layer of the communicating entities involved. Let e_1 and e_2 be two communicating entities and f an application defining an interoperability mechanism. We distinguish several categories of such mechanisms: mechanisms applying between languages $f : l^{\text{L}}_{e_1} \rightarrow l^{\text{L}}_{e_2}$, corresponding to a **dictionary**; mechanisms applying on grammars $f : g^{\text{L}}_{e_1} \rightarrow g^{\text{L}}_{e_2}$, corresponding to **translator**. The literature contains other discrimination criteria, especially on translators [2]. The **level of interoperability** is defined as the percentage of domain elements to which it is possible

to associate a codomain element using the interoperation mechanism. If these mechanisms are bijective, interoperability is **complete**. Completeness impacts what operations can be achieved out through cooperation. If an inverse mechanism f^{-1} is implemented, interoperability is **reciprocal**. Reciprocity impact the possible direction of information exchange.

To enable an **information exchange** between two communicating entities, it is necessary for these 2 entities to be globally interoperable and for an information exchange to be implemented. Reciprocal interoperability is needed when bidirectional information exchange is needed. To understand what is the state of the implementation of information exchange, a quantitative assesment is required. In the next section, we use interoperability model to propose a quantitative metric for assessing the state of information exchange in Open Science.

3.3 Open Science information exchange quantitative assessment

To assess the need for a solution and the type of solution needed (structural, incentive, ...) [16], this assessment is required. To the best of our knowledge, there is no quantitative evaluation available on the information exchange in Open Science. We propose to assess this quantity of information exchange through the percentage of data management platforms that exchange information in Open Science. We use data available on Re3Data, a catalog of data management platforms. This catalog contains information on 3117 research data management platforms (as of May 31, 2023) and is used by the European Union as an indicator of the state of open research[6]. We assume that an evaluation based on this catalog will provide a close-to-reality view of Open Science state. In the context of data exchange between Open Science platforms, two kinds of heterogeneity may prevent the platforms from being interoperable: the type of API provided by each platform to access to datasets (technical interoperability) and the kind and structure of the metadata used to describe the datasets (semantic interoperability). Metadata can be represented thanks to various schemas, vocabularies, thesauri and/or ontologies, that we will refer to as "metadata models" in the following.

Among the APIs used by the platforms described in Re3data, the only one that natively integrates an information exchange mechanism is OAI-PMH. This protocol enables metadata to be harvested on queried platform and to access the platform data from an external application, a catalog or another platform. However, OAI-PMH has a limitation when it comes to scaling up. Harvesting must be carried out by the harvester or the service provider, but it is necessary for these platforms to know each other beforehand. With the large number of existing platforms, it is not a viable approach. For the rest, we will assume that all platforms know each other, creating an overestimation of information exchange possibilities thanks to OAI-PMH. This hypothesis compensates for the lack of information on real exchanges between platforms.

[6] `https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en`

To represent the state of Open Science, we define the undirected graph of information exchange in Open Science

$$G_{OSci} = (V_{OSci}, E_{OSci})$$

with V_{OSci} the set of vertices in the graph, where each vertex is a data management platform, and E_{OSci} the set of edges, where each edge defines an information exchange capability between 2 platforms. A platform $v \in V_{OSci}$ is defined with 2 components $v = (API, MD)$, based on the 2 heterogeneity issues previously mentioned. An edge exists between 2 platforms if both implement OAI-PMH and have at least 1 metadata model in common. The visualization of the graph in Fig. 2 clearly illustrates the problem of lack of information exchange in Open Science. A large number of data management platforms (2827 nodes, $\sim$ 90%), materialized as grey spots, are unable to exchange their data with other platform. A set of 290 interconnected nodes is visible ($\sim<$ 10% of

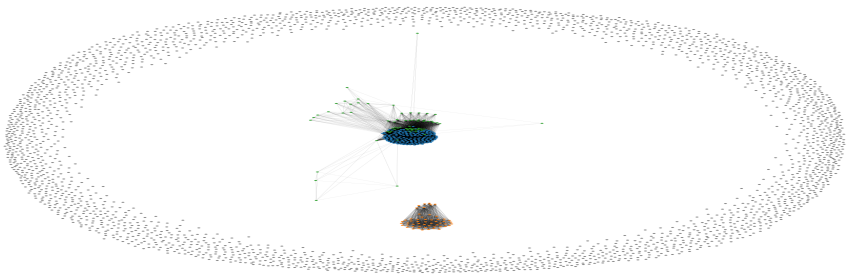


Fig. 2: Open Science information exchange graph visualization

total platforms), which materialize the platforms implementing OAI-PMH. Distinct communities can be observed based on node colors (automatically extracted using the Louvain method) with 2 distinct connected components, giving an indication of the lack of homogeneity in communities of Open Science regarding their ability to exchange information. The density of the graph is ~ 0.0046, close to a set of unconnected nodes. This level of information exchange appears very low through this visualization. For a more precise interpretation, we define the event X_d "Find the desired data d on a platform". This event occurs when the search is carried out on a platform that provides access to its data. The empirical probability $Pr(X_d)$ is equal to the number of platforms on which a dataset is available $|avail(d)|$, with $avail(d)$ the function that returns the set of platforms where d is found, divided by the total number of platforms in Open Science $|V_{OSci}|$.

$$Pr(X_d) = \frac{|avail(d)|}{|V_{OSci}|}$$

OAI-PMH does not include a mechanism for harvesting data from indirect neighbours. We define h, a number of hops between two platforms. From a network

point of view, counting hops may differ according to the protocols. In our context, we define the number of hops between 2 nodes as the distance between the two nodes on the path studied, i.e. the number of edges between these 2 nodes.

The maximum hop in OAI-PMH is 1. $avail(d)$ becomes $|\bigcup_{n=0}^{h} S^n(avail(d)|$, with $S : \mathbb{N} * \mathcal{P}(V_{OSci}) \rightarrow \mathcal{P}(V_{OSci})$ the successor function with hops such that $S^n(nodes) = \underbrace{S \circ ... \circ S(nodes)}_{\text{n times}}$, with $nodes \in \mathcal{P}(V_{OSci})$ and the successor function $S : \mathcal{P}(V_{OSci}) \rightarrow \mathcal{P}(V_{OSci})$ returning the set of neighbours of a set of nodes.

The probability of the event X_d when h hops are possible becomes

$$Pr^h(X_d) = \frac{|\bigcup_{n=0}^{h} S^n(avail(d)|}{|V_{OSci}|}$$

OAI-PMH allows to request platforms 1 hop away from the initial one. $avail(d)$ is equal to the average number of neighbours in Open Science plus the number of platforms managing that same dataset. We assume that data is not duplicated. The empirical probability of finding the desired data is equal to

$$Pr^1(X_d) = \frac{(1 + mean\_degree\_in\_OSci)}{total\_node\_number} \approx \frac{(1 + 14.24)}{3117} \approx 0.5\%$$

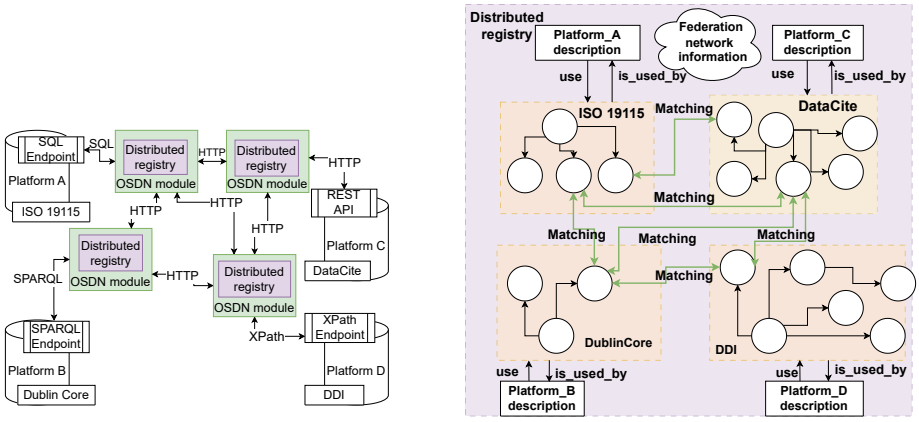
This probability is very low and confirms the description made by Open Science actors on the lack of cooperation between Open Science data management platforms and actors. This very low level of information exchange show the need to implement an information exchange in Open Science and not just improve it, **requiring architectural solutions** [16]. In the next section, we propose a network-based solution to implements information exchange in Open Science.

4 Proposition: The Open Science Data Network

Open Science contains many pre-existing data management platforms. Researchers are calling for decentralisation, unified research and control of open data by its owner. To interoperate existing systems and meet needs, we propose the Open Science Data Network (OSDN), a decentralised, federated and distributed interconnection network of data management platforms. We describe our solution in 2 parts. Firstly, we explore interoperability and information exchange within the OSDN. Then we explore a part of the scaling with the robustness of the solution, to ensure that the solution is sustainable.

4.1 Information exchange and interoperability in OSDN

This solution is based on a RESTful API using a registry shared among every platform (see Fig. 3a). This registry contains the information needed for the OSDN to operate (see Fig. 3b). A mechanism for propagating changes and



(a) Network architecture example (b) Distributed registry example

Fig. 3: OSDN implementation example

queries is implemented by broadcasting them to all neighbours until the changes or queries have been propagated to the whole network. From a technical point of view, this module takes the form of a Docker container with automatic deployment. The integration of this module requires a single operation: the implementation of the interoperation function between this module and the platform information retrieval mechanism, enabling the module to execute queries on the platform. **Reciprocal and complete technical interoperability** is achieved by standardising the exchange protocol between platforms based on the module. The registry contains information relating to the platforms (name, URL, interconnected platforms, etc.), the used metadata models (name, content, etc.) and the matchings between the metadata models.

We have observed a lack of adaptation of automatic matching solutions to the metadata models used in Open Science [7]. For the implementation of these matchings in OSDN, we propose a combined implementation of automatic and manual solutions. Firstly, the implementation of manual matchings is based on the establishment of community collaboration between the various actors in Open Science. Each platform manager sets up matches between its model and another model in the registry. Reducing the workload for each player goes some way to addressing the problem of scaling up manual matching. As it stands, the use of automatic matches can enable 2 solutions to be put in place: (i) a recommendation of matches reducing the cost of setting up manual matches, (ii) a crossing of the results of automatic matching solutions to improve the results benefiting from the collaborative structure of the OSDN. To address the problem of matching trustworthiness and the fact that they can contradict each other, we decided to keep all the matchings and associate them with a likelihood score, based on user feedback. The aim is to return query results based on the most likely matches. **Semantic interoperability** is achieved by setting

up gateways thanks to the matchings between models, and so implementing a **global interoperability**.

4.2 Scalability - Robustness of OSDN

Data management is a critical point in the research knowledge creation process. It is therefore necessary to ensure that the network is robust and durable, relating to theory of percolation [3]. The objective is to determine the number of nodes that need to be deleted in a network to allow the network to be split into several disconnected sub-networks. Deletions may be either *voluntary deletion*, i.e. carried out through network attacks, targeting nodes that can cause the most damage by deleting them; or *random deletions*, generally caused by a node's failure, disconnecting it from the network.

Choosing the right topology can provide particular resistance to these events, but it seems rather orthogonal to succeed in protecting against both voluntary and involuntary deletions [3]. A solution that minimizes vulnerability to these two events is found for a scale-free network topology with a single hub and with the other nodes all having a degree of five [3]. To reduce adoption costs, we set the minimum degree of node equal to two. This network topology follows a power distribution in the node degree distribution. In the following, we denote γ the power law parameter followed by the network topology [3]. This topology offers an interesting feature for information exchange. Based on epidemiological approach applied to networks [3], the propagation time of a pathogen τ (which can be interpreted as an information) in a scale-free networks with a γ less than 3 tends towards 0, when the number of nodes increases [3]. Moreover, scale-free network topologies following a power law with $2 < \gamma < 3$ are in ultra-small world regime [3]. These networks have average distance between nodes growth equal to $lnlnN$. Linking a large number of nodes create a small distance between them [3], leading to a small increase in information propagation time.

Inscription protocol: To be able to implement a specific topology, it is necessary to implement a control of the registration process. We chose a network parameter γ of 2.5 to avoid potential edge effects and to have the characteristics of networks with $2 < \gamma < 3$. This registration process is described by the algorithm 1.

The $get\_node\_nearest\_from\_distribution$ function returns a node of degree k_{opti} to connect to, such that k_{opti} minimizes the Kullback-Leibler divergence D_{KL}, giving a measure of dissimilarity between two distributions. In other words, it returns a degree k_{opti} node. We select the first node found in the list of degree k_{opti} nodes to reduce computation time in our implementation. To understand formally, let P be the initial degree distribution of the network graph with the node v_{chosen} added but not connected, P_n the degree distribution of the graph nodes after the connection of v_{chosen} to a node of degree n, Q a power law with $\gamma = 2.5$, E_k the set of nodes of degree k from the distribution P, and E the total set of nodes in the distribution P. We define that connecting v_{chosen} does not

Algorithm 1: Node inscription in network

Input : A network graph $G = (V, E)$, a node v_{add}, a node in the network to link to v_{chosen}
Output: A network graph $G = (V', E')$

if $|V| = 0$ **then**
  $add\_node\_to\_network(G, v_{add})$
end if
if $|V| = 1$ **then**
  $add\_node\_to\_network(G, v_{add})$
  $add\_edge(v_{add}, v_0)$ `/* ` v_0 `is the only vertice in the network */`
end if
if $|V| > 1$ **then**
  $add\_node\_to\_network(G, v_{add})$
  $add\_edge(v_{add}, v_{chosen})$
  $v = get\_node\_nearest\_from\_distribution(v_{add}, V)$ `/* Return the nearest node from set of nodes that should be connected to fit the most a power law with` $\gamma = 2.5$ `*/`
  $add\_edge(v_{add}, v)$
end if
return G

change its degree, assuming it has initially 2 self loops that we want to replace by connecting it to another node. We have $D_{KL}(P||Q) = \sum_{k \in K} q(k) \log(\frac{p(k)}{q(k)})$ with K the set of possible degrees in the graph, from 0 to k_{max} and $p(k)$ and $q(k)$, respectively the probability mass function of distribution P and Q. Connecting to a node of degree n increases the degree of the target node by 1. So we remove 1 node of degree n and add a node of degree $n+1$ to the distribution. We have

$$\begin{cases} p_n(k) = \frac{|E_k|-1}{|E|} = \frac{|E_n|-1}{|E|} \text{ for } k = n \\ p_n(k) = \frac{|E_k|+1}{|E|} = \frac{|E_{n+1}|+1}{|E|} \text{ for } k = n+1 \\ p_n(k) = P(k) = \frac{|E_k|}{|E|} \text{ otherwise} \end{cases}$$

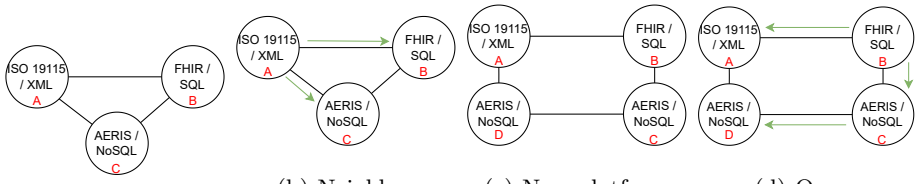
where $p_n(k)$ is the empirical mass function of the distribution P_n. $p_n(k)$ is independent of n when $k \notin \{n, n+1\}$. We end up with the optimization problem in (1). We have $q(n) = \frac{n^{-\gamma}}{\zeta(\gamma)} = \frac{n^{-2.5}}{\zeta(2.5)}$, with ζ the Riemann zeta function [3]. Since the value of $\zeta(2.5)$ is independent of n, equation (1) can be simplified into (2).

$$k_{opti} = \arg\min_n (p_n(n) \log(\frac{p_n(n)}{q(n)}) + p_n(n+1) \log(\frac{p_n(n+1)}{q(n+1)})) \quad (1)$$

$$k_{opti} = \arg\min_n (\frac{|E_n| - 1}{|E|} \log(\frac{\frac{|E_n|-1}{|E|}}{n^{-2.5}}) + \frac{|E_{n+1}| + 1}{|E|} \log(\frac{\frac{|E_{n+1}|+1}{|E|}}{(n+1)^{-2.5}})) \quad (2)$$

Equation (2) gives the optimal degree of the node to be computed in the *get_node_nearest_from_distribution* function. Knowing that we are evaluating n

from 2 to k_{max}, with k_{max} the highest degree in the distribution before adding a node, the registration function has a computational complexity in $O(k_{max})$.



(a) Initial setup (b) Neighbor querying (c) New platform inscription (d) Query propagation

Fig. 4: POC implementation

5 Experiments

We experimented our solution in 2 distincts parts, based on the development of a proof of concept for OSDN for an assessment of our solution's ability to integrate platforms and metadata models from Open Science and then we developped a supplementary Web GUI to make an user experiment with agronomy researcher, in order to observe the benefits. All code is open, accessible and re-executable.

5.1 OSDN Network POC

Use case / Domain	Information retrieval	Data integration	Core data set	Secondary use
Humanities	**(14)**: (Struct) (SDMX) SDMX - Statistical Data and Metadata Exchange	**(11)**: (Struct) (OAI) OLAC	**(4)**: (Struct) (DDI) DDI - Data Documentation Initiative Metadata Standard	**(16)**: (Struct) (TEI) Text Encoding Initiative Guidelines
Life science	(Sem) (WHO) ICD-10	**(3)**: (Struct) (TDWG) Darwin Core		**(8)**: (Struct) (HL7) FHIR **(2)**: (Struct) (HL7) C-CDA
Natural science	**(12)**: (Struct) PDB		**(9)**: (Struct) (ISO) ISO-19115 **(1)**: (Struct) AERIS	
Engeeniering	**(7)**: (Struct) (RDA) EngMeta – Metadata for Computational Engineering		**(15)**: (Struct) (OGC) SensorML **(17)**: (Struct) (OGC) CoverageJSON	
General	(Sem) (ISO) ISO 639-2 **(6)**: (Struct) e-Government Metadata Standard	**(13)**: (Struct) (OCLC/RLG) PREMIS: Data Dictionary for Metadata Preservation	**(5)**: (Struct) Dublin Core **(10)**: (Struct) (DataCite) DataCite	

Table 1: Classification of metadata models

To assess the technical feasibility of OSDN and its mechanisms, we have developed a Proof of Concept[7]. We have integrated 3 platforms from Open Science with different metadata management technologies, with MySQL, MongoDB and an XML-based solution (see Fig. 4a). We executed a query on one platform and

[7] `https://github.com/vincentnam/Openscience_network_experiment`

verified the propagation of this query to its neighbours (see Fig. 4b). Then, we evaluated the query propagation to indirect neighbours by adding a new node (Fig. 4c) and re-executing the query that also returns the dataset information from the new platform (Fig. 4d). To ensure that OSDN can integrate the metadata models from Open Science, we integrated and interoperate 19 metadata models in OSDN registry. To avoid biases, we selected them after a systematic review of metadata models [26] according to 3 criteria: domain, use case and type (structural (17) or semantic (2)). We added the model creator as supplementary criteria (see Table 1). We manually created matches between structural models to validate that models can be interoperated (see Table 2).

Concept \ Model number (see Table 1)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Dataset Title	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Content localization	x		x					x		x							x
Content UOM	x	x					x								x		

Table 2: Models interoperability : models matching

5.2 Use case - An agronomic research project

To estimate the benefits of our solution, we developed a graphical interface to enable graphical information retrieval from a set of 11 data management platforms from different domains and communities[8], from which we downloaded a part of the data catalog. A single platform allowed us to download the whole catalog. We worked with Dr. Thomas Pressecq[0000−0003−0067−7903]. During the last 3 years, his research focused on the development of a decision support system to guide the use of biocontrol tools by farmers [18]. The main problem he met was the lack of data. Dr. Pressecq described a lateness in agronomy data management, with too few data and too few data management platforms either used or known by researchers. To carry out his research project, Dr. Pressecq extracted 381 row in his dataset, from 900 scientific publications on 41 strains of micro-organisms. Each row contains the combination of a biocontrol agent and its associated efficacy factor (propriety of biocontrol agent, environmental conditions, cropping practices, propriety of the pathogen). He estimates that processing a scientific article requires 20 minutes, including reading, analysing and extracting information. We asked him to reproduce his search for data, about a sub-part of his dataset (on "trichoderma harzianum T-22" strain) on OSDN. On this specific strain, his dataset contains 29 tabular rows extracted from a total of 115 scientific publications, for a mean time by line of 79 minutes. He performed a simple query applying only to the title or the description. We evaluated the time reduction gained using OSDN compared with his past experience to build the dataset. As side effect benefits, using OSDN allowed to get more datasets,

[8] `https://github.com/vincentnam/RCIS_userfeedback_experiment`

to increase the volume of collected data and to generate new perspectives for future works.

TIME: Dr. Pressecq kept 17 datasets or publications from 3 different platforms out of the 11 platforms (13 results on NCBI[9], 3 on the Harvard generalist dataverse[10] and 2 on Figshare, a generalist platform[11]). The total time spent on this search was 2 hours and 30 minutes. Of the 17 results selected, 12 were new datasets initially unknown to Dr. Pressecq. However, of the 29 tabular rows in the dataset, 3 rows could have been replaced by datasets containing extracted and usable data. With our solution, it took 45 minutes to Dr. Pressecq to find 3 new lines, which corresponds to a reduction in dataset construction time on this 3 lines of $\sim 80\%$ (compared to 3 times 79 minutes) thanks to datasets reuse.

DATASET ENRICHMENT: OSDN provided 2 new lines that were not found by the initial method. This represents a **7%** increase in the dataset volume.

INTERDISCIPLINARITY: OSDN provided datasets from several domains (bioinformatics and health (NCBI), social science (Harvard dataverse)) and other communities (Figshare and Harvard). It shows an interdisciplinary and intercommunity data enrichment. But, this interdisciplinary and intercommunities information exchange also provided datasets that allow the creation of new futures works. Dr. Pressecq described several new futures works which we divide into 2 categories :

- **Intradisciplinary research works** with a study based on data crossing of soil characteristics (FORM@TER[12]) for microbial agents prospection in different types of soil.
- **Interdisciplinary research works**, by crossing data taken from the European data platform[13], to assess the perception of biocontrol tools among farmers and consumers at European level as a collaboration of agronomy and social sciences domains.

6 Conclusion

Open Science still faces major obstacles to sharing and finding data, like the lack of coordinated solutions between data platforms [8]. We have proposed a quantitative assessment of the current state of information exchange in Open Science, based on percentage of platform that exchange information, estimated at 0.5%. To answer this lack of information exchange implementation in Open Science, we have proposed OSDN, a decentralized, distributed and federated network solution for research data management platforms. We assessed the technical feasibility and behaviour of this network. Finally, we identified the contributions of OSDN in the context of a real research project, saving time in building datasets, data enrichment and datasets reuses and research projects, and providing new interdisciplinary research perspectives. We observed that our solution allows information exchange. But this solution faces a real challenge : adoption. Even if it

[9] www.ncbi.nlm.nih.gov [10] dataverse.harvard.edu [11] figshare.com
[12] https://www.poleterresolide.fr/ [13] data.europa.eu

has be thought to be as easy as possible to integrate OSDN, the development cost of the interoperability function of the OSDN module and the inscription process cost, especially model matching, may vary due to different contexts of data management platform (human resources, technologies used, etc..). Moreover, it may lead to an additional burden to process network queries. This could discourage platform manager without proper incentives. As a futures works, in addition to working on the adopting cost of this solution, there a 2 different axes. Firstly, an exploration of semantic, semantic interoperability concept and the problematic of automatic matching algorithms in the context of Open Science will be explored. Then, we plan to explore the scalability and several optimisation on resources consumption in the OSDN.

Aknowledgment

We thank Dr. Thomas Pressecq from INRAE (Institut national de la recherche agronomique) and APREL (Association Provençale de Recherche & d'Exprimentation Légumière) for his participation in our experiment and his feedback.

References

1. Ackoff, R.L.: From data to wisdom. J. of applied syst. analysis **16**(1), 3–9 (1989)
2. Aho, A.V., Hopcroft, J.E., Ullman, J.D.: A general theory of translation. Mathematical Systems Theory **3**, 193–221 (1969)
3. Barabási, A.L., Pósfai, M.: Network science. Cambridge University Press, Cambridge (2016), http://barabasi.com/networksciencebook/
4. Beyan, O., et al.: Distributed analytics on sensitive medical data: the personal health train. Data Intelligence **2**(1-2), 96–107 (2020)
5. Corpas, M., et al.: A fair guide for data providers to maximise sharing of human genomic data. PLoS computational biology **14**(3), e1005873 (2018)
6. Costin, A., Eastman, C.: Need for interoperability to enable seamless information exchanges in smart and sustainable urban systems. Journal of Computing in Civil Engineering **33**(3), 04019008 (2019)
7. 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)
8. Digital Science, Hahnel, M., Smith, G., henning schoenenberger, Scaplehorn, N., Day, L.: The State of Open Data 2023 (11 2023). https://doi.org/10.6084/m9.figshare.24428194.v1
9. Guizzardi, G.: On ontology, ontologies, conceptualizations, modeling languages and (meta) models. Databases and Information Systems IV, IOS Press (2007)
10. Heiler, S.: Semantic interoperability. ACM CSUR **27**(2), 271–273 (1995)
11. Hughes, L.D., et al.: Addressing barriers in fair data practices for biomedical data. Scientific Data **10**(1), 98 (2023)
12. Jacobsen, A., et al.: Fair principles: interpretations and implementation considerations. Data intelligence **2**(1-2), 10–29 (2020)
13. Kathawalla, U.K., Silverstein, P., Syed, M.: Easing into open science: A guide for graduate students and their advisors. Collabra: Psychology **7**(1), 18684 (2021)

14. Litwin, W., Mark, L., Roussopoulos, N.: Interoperability of multiple autonomous databases. ACM Computing Surveys (CSUR) **22**(3), 267–293 (1990)
15. National Academies of Sciences and Global Affairs and Board on Research Data and Information and Committee on Toward an Open Science Enterprise: Open science by design: Realizing a vision for 21st century research (2018)
16. Nosek, B.: Strategy for culture change. Center for Open Science **11** (2019)
17. Noura, M., et al.: Interoperability in internet of things: Taxonomies and open challenges. Mobile networks and applications **24**(3), 796–809 (2019)
18. Pressecq, T., et al.: Decicontrol: a participative tool to gather & share data regarding biocontrol efficacy. In: European Scientific Conference–Towards Pesticide Free Agriculture "What research to meet the pesticides reduction objectives embedded in the European Green Deal?" (2022)
19. Rainey, L., Lutomski, J.E., Broeders, M.J.: Fair data sharing: An international perspective on why medical researchers are lagging behind. Big Data & Society **10**(1), 20539517231171052 (2023)
20. Sadeh, Y., et al.: Opportunities for improving data sharing and fair data practices to advance global mental health. Cambridge Prisms: Global Mental Health **10**, e14 (2023)
21. Spies, J.R.: The open science framework: Improving science by making it open and accessible. University of Virginia (2013)
22. Tanhua, T., et al.: Ocean fair data services. Frontiers in Marine Sc. **6**, 440 (2019)
23. Thomesse, J.: Interoperability: an overview. IFAC Proc. Vol. **30**(7), 433–438 (1997)
24. Tolk, A., et al.: Applying the levels of conceptual interoperability model in support of integratability, interoperability, and composability for system-of-systems engineering. Journal of Systems, Cybernetics, and Informatics **5**(5) (2007)
25. Top, J., et al.: Cultivating fair principles for agri-food data. Computers and Electronics in Agriculture **196**, 106909 (2022)
26. Ulrich, H., et al.: Understanding the nature of metadata: systematic review. Journal of medical Internet research **24**(1), e25440 (2022)
27. Uribe, S.E., et al.: Dental research data availability and quality according to the fair principles. Journal of dental research **101**(11), 1307–1313 (2022)
28. Uschold, M.: Where are the semantics in the semantic web? AI Magazine **24**(3), 25–25 (2003)
29. Van Der Veer, H., Wiles, A.: Achieving technical interoperability. European telecommunications standards institute (2008)
30. Vesteghem, C., et al.: Implementing the fair data principles in precision oncology: review of supporting initiatives. Briefings in bioinformatics **21**(3), 936–945 (2020)
31. Vicente-Saez, R., Martinez-Fuentes, C.: Open science now: A systematic literature review for an integrated definition. Journal of business research **88**, 428–436 (2018)
32. Wegner, P.: Interoperability. ACM CSUR) **28**(1), 285–287 (1996)
33. Wileden, J.C.o.: Specification-level interoperability. Communications of the ACM **34**(5), 72–87 (1991)
34. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
35. Zeng, M.L.: Interoperability. KO Knowledge Organization **46**(2), 122–146 (2019)