Recherche de données interdisciplinaire dans la Science Ouverte

Vincent-Nam Dang

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse Capitole (UT Capitole)*

Présentée et soutenue le *25 Octobre 2024* **par :**

Vincent-Nam DANG

Recherche de données interdisciplinaire dans la Science Ouverte

JURY

Anne LAURENT	Professeure, Université de Montpellier	Rapporteure
Cécile FAVRE	Professeure, Université Lyon 2	Rapporteure
Cyril LABBE	Professeur, Université Grenoble Alpes	Examinateur
Imen MEGDICHE	Maîtresse de Conférences, INU Jean-François-Champollion	Encadrante
Nathalie AUSSENAC-GILLES	Directrice de Recherche, IRIT, CNRS	Co-Directrice de thèse
Franck RAVAT	Professeur, Université Toulouse Capitole	Co-Directeur de thèse

École doctorale et spécialité :
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse : Informatique et Télécommunications

Unité de Recherche :
IRIT: Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :
Franck RAVAT et *Nathalie AUSSENAC-GILLES*

Rapporteures :
Anne LAURENT et *Cécile FAVRE*

Remerciements

Ce travail de recherche de trois ans n'aurait pas été possible sans les efforts de toutes les personnes ayant participé à ma vie durant cette période, m'ayant épaulé et écouté. Je souhaite remercier plusieurs personnes spécifiquement mais je n'oublie pas l'ensemble de celles ayant eu un impact direct ou indirect sur cette thèse.

Tout d'abord, je souhaiterais exprimer ma gratitude et ma reconnaissance à mes directeurs de thèse. Je souhaite remercier le Pr. Franck RAVAT qui m'a épaulé et guidé tout au long de ce voyage dans le monde de la recherche et a permis que cette thèse voie le jour. Sa disponibilité et sa rigueur m'ont conduit à grandement améliorer mes travaux dans toutes leurs facettes. Je souhaite ensuite remercier Nathalie AUSSENAC-GILLES qui m'a poussé dans mes retranchements, me permettant de percevoir et d'approfondir des perspectives nouvelles. Je remercie Imen MEGDICHE, encadrante de thèse, qui m'a apporté un accompagnement dans la production de mes travaux, sans qui cette thèse n'aurait pas été la même.

J'adresse tous mes remerciements aux membres du jury, Pr. Anne LAURENT, Pr. Cécile FAVRE et Pr. Cyril LABBE, qui ont accepté d'évaluer mon travail et m'ont permis de soutenir ma thèse malgré la charge de travail associée et un emploi du temps rempli durant cette période de rentrée universitaire. Je souhaite aussi remercier particulièrement les rapporteures de cette thèse, Pr. Anne Laurent et Pr. Cécile FAVRE, pour leurs commentaires et avis précieux sur mes travaux.

Je remercie l'université Toulouse Capitole ainsi que la région Occitanie, au travers du projet SO :LDI, d'avoir financé cette thèse.

Je souhaite aussi remercier tous les enseignants-chercheurs de l'IRIT que j'ai pu rencontrer durant tout mon parcours universitaire, dont Pr. Karen PINEL-SAUVAGNAT, Pr. Rahim KACIMI, Dr. François THIEBOLT, Pr. Marie-Pierre GLEIZE et toute l'équipe du projet neOCampus, sans qui je n'aurais pu en arriver jusqu'ici. Toutes ces personnes m'ont aiguillé, conseillé et guidé tout au long de mon parcours.

Enfin, je tiens aussi à remercier tous mes amis et mes collègues du bureau des doctorants, toute ma famille pour leur soutien et leur aide dans ce voyage, mes frères et soeur, mes neveux et nièces qui m'ont mis du baume au coeur dans les périodes de doute. Je tiens particulièrement à remercier ma mère, qui a toujours été là pour moi, qui m'a toujours poussé et encouragé.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte

La Science Ouverte est définie comme “un accès ouvert aux connaissances partagées et développées à travers un réseau de collaborations dans le monde de la recherche scientifique” (Vicente-Saez and Martinez-Fuentes (2018)). Le monde de la recherche scientifique est composé de nombreuses communautés.

Chaque communauté nécessite le déploiement d’outils adaptés aux besoins intracommunautaires. Ces communautés de chercheurs sont organisées autour de points de rassemblement communautaire (cf. Figure 1.1) pouvant être un projet de recherche, une organisation, des outils, une discipline, etc...

Une classification des différentes disciplines de la recherche est réalisée par l’organisme de recherche allemand, la DFG (“Deutsche Forschungsgemeinschaft”) [1]. Ces disciplines scientifiques, au nombre de quatre (“Humanités et Sciences Sociales”, “Sciences de l’Ingénierie”, “Sciences de la Vie” et “Sciences Naturelles”) sont décomposées en sous-niveaux, sous forme d’un arbre (cf. Figure 1.2), jusqu’à atteindre les sujets de recherche. La DFG recense deux cent soixante-quinze sujets de recherche différents dans les quatre disciplines de recherche. Chaque sujet de recherche et discipline possède des types de données et des besoins spécifiques.

La **recherche interdisciplinaire** est vue comme un atout, permettant de répondre à des questions de recherche trop difficiles à traiter sous l’angle d’une seule discipline (Corbett et al. (2013)). Cette interdisciplinarité est de plus en plus présente au cœur des questions de recherche abordées par les chercheurs (Ramachandran et al. (2021)). Une partie de l’interdisciplinarité et de son déploiement passe par la recherche de données de recherche (jeux de données et publications scientifiques) provenant de différentes disciplines et communautés, dont le croisement permet d’apporter de nouveaux points de vue sur les questions de recherche et de faire émerger de nouvelles questions de recherche.

Les données sont disséminées dans un très grand nombre de plateformes (Tanhua et al. (2019)). Ce très grand nombre de plateformes permet la gestion d’un très grand volume de données de recherche. Ces données de recherche sont potentiellement des données de recherche de haute valeur (Tanhua et al. (2019)) et permettre leur utilisation offrirait un fort enrichissement des processus de recherche scientifique.

Les besoins de chaque communauté et chaque discipline engendrent une haute variété dans ces plateformes. Cette variété se traduit par des différences de technologies uti-

1. `https://www.dfg.de/en/research-funding/proposal-funding-process/interdisciplinarity/subject-area-structure`

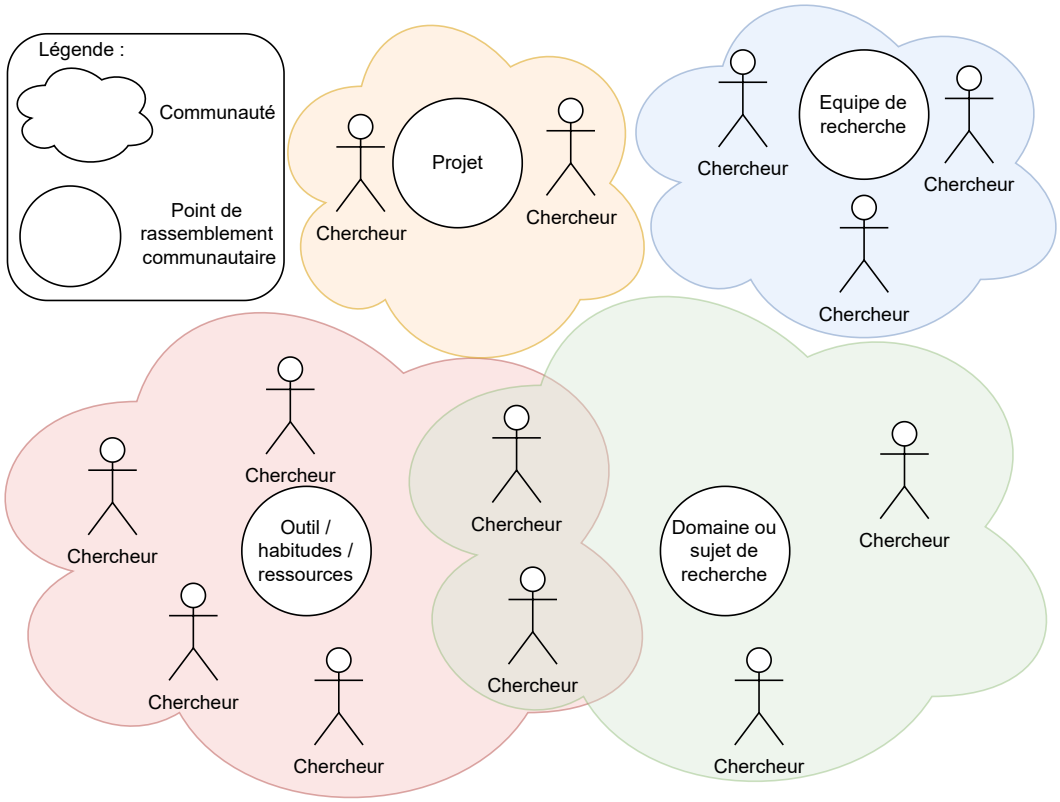

FIGURE 1.1 – Les communautés dans la recherche

lisées pour gérer les données mais aussi des différences de modélisation des métadonnées nécessaires à la recherche de ces données. En 2016, Wilkinson et al. (2016) mentionnait déjà l'existence de plus de 600 standards de définition de métadonnées inscrits sur le site BioSharing (devenu Fairsharing [2]), avec des modèles, des vocabulaires et des recommandations. En 8 ans, ce chiffre a presque triplé pour atteindre 1774 standards différents (visité le 15 Août 2024).

1.2 Problématique scientifique

Prenons l'exemple d'une entité de recherche (un chercheur, une équipe de recherche ou un laboratoire de recherche) nommée ER ("Entité de Recherche"). ER fait partie d'une communauté de chercheurs en astrobiologie et souhaite trouver une réponse à une question de recherche. Cette communauté est hautement interdisciplinaire (Poisot et al. (2019)). ER connaît cinq plateformes utilisées dans sa communauté. Ces cinq plateformes permettent à ER de réaliser une recherche interdisciplinaire de données nécessaire à l'élaboration d'une réponse à sa question de recherche. Cette recherche doit être réalisée sur l'ensemble de ces plateformes. Cependant la recherche de données est décrite comme trop difficile à réaliser pour les chercheurs (Poisot et al. (2019)) et le coût d'apprentissage de l'usage

2. `http://fairsharing.org/`

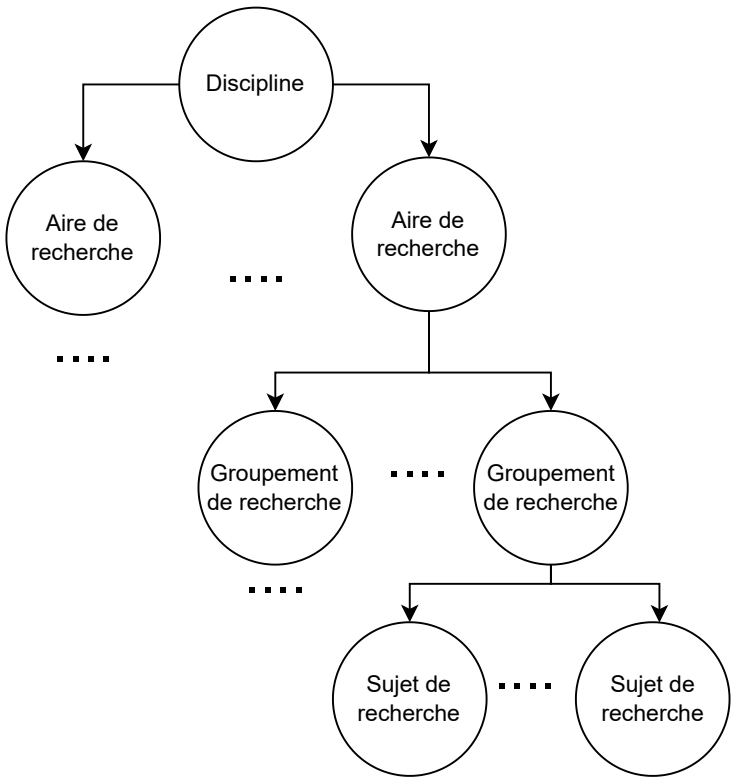

FIGURE 1.2 – Structure de la classification des sujets de recherche de la DFG

des plateformes de données est considéré comme trop élevé (Van Panhuis et al. (2014); Top et al. (2022)). De plus ER ne connaît qu'une partie des plateformes utilisées dans sa communauté, l'empêchant de trouver des données potentiellement utiles à son projet de recherche.

Problématique scientifique

Dans le contexte de la Science Ouverte, comment mettre en place une exploration multi-plateformes pour répondre à des objectifs de recherche interdisciplinaire de données ?

Cette question de recherche nécessite la réponse à un verrou :

Verrou scientifique

Être capable d'accéder à une grande variété de plateformes qui implique d'être capable de gérer des données de recherche (jeux de données et publications scientifiques)

Cet accès à une variété de plateformes implique d'être capable de :

Décomposition du verrou scientifique

1 • Définir un accès multi-plateformes et multi-modèles pour réaliser une recherche de données ;
2 • Proposer une solution au grand volume de données gérées par ces plateformes ;
3 • Proposer une solution permettant le passage à l'échelle de cette recherche de données.

1.3 Objectif des travaux de recherche

Cette thèse a pour objectif de proposer une solution d'exploration multi-plateformes afin de permettre une recherche de données interdisciplinaire dans le contexte de la Science Ouverte. Pour chaque verrou, nous avons pour objectif de :

1 • Définir les besoins et les contraintes qui s'attachent à la définition de cette exploration multi-plateformes et multi-modèles ;
2 • Proposer une solution intra-communautaire de recherche unifiée sachant le grand volume de données de recherche ;
3 • Proposer une solution pour cette recherche unifiée et interdisciplinaire de données, en prenant en compte le passage à l'échelle de cette solution.

1.4 Plan du manuscrit

Dans le chapitre 2, nous proposons une exploration du concept d'interopérabilité, nécessaire à la définition de cette exploration multi-plateformes et multi-modèles. Nous réalisons une étude comparative de onze travaux provenant de disciplines et de communautés différentes. Ensuite, nous proposons une exploration des différentes solutions de gestion, d'accès ou de recherche de données ouvertes dans la Science Ouverte, que nous appelons génériquement "Plateforme de Données de Recherche Ouverte" (PDRO).

Dans le chapitre 3, nous proposons une compréhension générique et exhaustive de l'interopérabilité, permettant de répondre aux lacunes observées dans la littérature. Cette proposition nous sert à définir la problématique et les verrous de l'interopérabilité des PDRO en vue de mettre en place une recherche de données unifiée. En se basant sur notre proposition, nous proposons une analyse quantitative de l'échange de métadonnées dans la Science Ouverte, nécessaire à une recherche unifiée de données.

Dans le chapitre 4, nous proposons une extension du concept de lac de données pour permettre une recherche et une consommation de données intracommunautaire. Ce Lac de Données de la Science Ouverte (LDSO) permet d'ingérer de nouvelles sources de données et de nouveaux profils d'utilisateurs. Cette proposition est composée d'une définition du concept de LDSO, d'une architecture fonctionnelle et d'une architecture technique permettant son implantation.

Dans le chapitre 5, nous proposons un réseau décentralisé, distribué et fédéré pour mettre en place une recherche unifiée, interdisciplinaire et intercommunautaire de données de recherche. Ce Réseau de Données de la Science Ouverte (RDSO) est pensé pour être robuste, facile à intégrer et permettant cette recherche de données sans charge supplémentaire pour les chercheurs.

Dans le chapitre 6, nous explicitons le développement de preuves de concepts pour le LDSO et le RDSO et nous présentons des expériences utilisateurs avec des chercheurs. Ces expérimentations nous permettent de valider les solutions proposées et mettent en avant les avantages de celles-ci au travers d'une évaluation quantitative.

Dans le chapitre 7, nous faisons un bilan sur nos différentes propositions permettant d'implanter un réseau de collaborations entre chercheurs dans le contexte de la Science Ouverte. Nos solutions apportent aux chercheurs une réduction du coût de la recherche de données et un enrichissement du volume de données trouvées. Nous complétons ce bilan par la proposition d'activités de recherche complémentaires à court, moyen et long terme.

Chapitre 2

Etat de l'art

La **recherche interdisciplinaire** est vue comme un atout, permettant de répondre à des questions de recherche trop difficiles à traiter sous l'angle d'une seule discipline (Corbett et al. (2013)). Cette interdisciplinarité est de plus en plus présente au cœur des questions de recherche abordées par les chercheurs (Ramachandran et al. (2021)). Une partie de l'interdisciplinarité et de son déploiement passe par la recherche de données provenant de différentes disciplines, dont le croisement innovant permet d'aborder de nouvelles questions de recherche. Ainsi, la recherche et le partage de données de recherche sont essentiels au bon fonctionnement du processus de création de connaissances au sein et entre les communautés.

Dans la pratique, les données de recherche sont échangées entre chercheurs proches et sortent rarement du cadre des équipes de recherche ou des laboratoires, ce qui empêche les autres chercheurs de la communauté d'accéder à ces données (Aydinoglu et al. (2014)).

La recherche de données est perçue comme trop difficile à réaliser pour les chercheurs (Poisot et al. (2019)) et le coût d'apprentissage de l'usage des plateformes de données est considéré comme trop élevé (Van Panhuis et al. (2014); Top et al. (2022)). Pour résoudre ce problème, les chercheurs demandent un accès unifié aux métadonnées décrivant les jeux de données (Tanhua et al. (2019)) et souhaitent une recherche d'informations transparente sur toutes les plateformes de données de recherche.

L'interopérabilité est la solution pour permettre la mise en place de cet accès unifié (Ise (2014)). Pour proposer une solution d'accès unifié aux métadonnées, il est nécessaire de comprendre ce concept d'interopérabilité et surtout son application aux plateformes de données de recherche ouvertes de la Science Ouverte. Ainsi dans la section 2.1, nous analysons différentes propositions d'interopérabilité pour comprendre ce concept ainsi que les verrous qui sont posés par le besoin d'interopérabilité des plateformes de données de recherche ouvertes. Nous identifions aussi ses différentes composantes. Dans la section 2.2, nous nous centrons sur la Science Ouverte en définissant le concept de Plateformes de Données de Recherche Ouvertes (**PDRO**) en proposant une analyse comparative de celles-ci selon les composantes de l'interopérabilité présentées dans la section précédente.

2.1 L'interopérabilité

L'interopérabilité est une caractéristique transverse présente dans de nombreux domaines. La définition de ce concept varie selon les domaines et les besoins exprimés. Or, la Science Ouverte repose sur de très nombreuses plateformes, avec des données hétérogènes et décrites par des métadonnées modélisées par de nombreux modèles différents. Un des

Sujet/domaine	U.E.	Général	IoT	K.O	FAIR	Télécom.	Standard MD	Entreprise			Cloud
Type d'interopérabilité	1	2	3	4	5	6	7	8	9	10	11
Selon les entités à rendre interopérables											
Des entreprises									x	x	
Des clouds											x
Des appareils			x								
Des données									x	x	
Des sociétés										x	
Des informations										x	
Des réseaux			x					x			
Des plateformes			x								
Des processus										x	
Des services										x	
Des profiles							x				
Des savoirs									x		
Des communications									x		
Des plateformes IaaS/PaaS/SaaS											x
Des plateforme de cloud											x
Des gest. de plat. de cloud											x
Des communautés											
Des applications									x		
Selon la caractérisation de l'interopérabilité											
Technique	x	x				x		x			
Syntaxique		x	x	x		x	x	x			
Sémantique	x	x	x	x		x		x	x		
Sémantique formelle							x				
Pragmatique		x						x			
Dynamique		x									
Conceptuelle		x									
Structurelle				x							
Informationnelle								x			
Internationnale											
Légale	x										
Organisationnelle	x					x		x			
Système				x							

TABLE 2.1 – Décomposition du problème d'interopérabilité en types d'interopérabilité par les travaux étudiés ; Légende des travaux - 1 : Corcho et al. (2021) 2 : Tolk et al. (2007) 3 : Noura et al. (2019) 4 : Zeng (2019) 5 : Wilkinson et al. (2016) 6 : Van Der Veer and Wiles (2008) 7 : Nilsson et al. (2008) 8 : Ambrosio and Widergren (2007) 9 : Zwegers (2003) 10 : Berre et al. (2007) 11 : Kostoska et al. (2016)

objectifs de cette thèse est donc de proposer une solution complète et exhaustive pour l'interopérabilité des différentes plateformes de la Science Ouverte. Dans cette section, nous étudions les travaux relatifs à l'interopérabilité présents dans la littérature. Les critères que nous avons retenus pour analyser ces travaux sont les suivants :

— les **entités** à rendre interopérables ;
— les éventuelles **démarches d'implantation** ;
— les **mécanismes d'implantation** ;
— les différents **types d'interopérabilité** à mettre en place pour gérer la complexité du concept d'interopérabilité.

Il manque une compréhension commune intégrant tous ces critères pour l'appliquer au besoin de partage de données dans la Science Ouverte (Rezaei et al. (2014a)). Nous proposons une analyse des travaux existants sur l'interopérabilité pour observer les différences et les points communs. Dans la section 2.1.1, nous présentons différents travaux ainsi que nos critères de sélection. Dans les sections suivantes, nous analysons ces travaux selon différents paramètres : (i) les types d'entités à rendre interopérables (cf. Section 2.1.2)

(ii) les mécanismes implantés pour assurer cette interopérabilité, (cf. Section 2.1.3), (iii) la démarche d'implantation (cf. Section 2.1.4), (iv) la décompositions de l'interopérabilité en types d'interopérabilité (cf. Section 2.1.5).

2.1.1 Les travaux étudiés sur l'interopérabilité

Nous avons sélectionné 11 travaux relatifs à la notion d'interopérabilité :

— l'European Open Science Cloud Interoperability Framework (Corcho et al. (2021)), basé sur l'European Interoperability Framework (EIF), correspond à la mise en place de services publics européens. Cette proposition s'appuie sur la définition suivante de l'interopérabilité : "capacité des organisations à interagir en vue d'objectifs mutuellement bénéfiques, impliquant le partage d'informations et de connaissances entre ces organisations, par le biais des processus d'entreprises qu'elles soutiennent, grâce à l'échange de données entre leurs systèmes d'information et de communication". Cette définition, proposée par l'EIF, décompose le problème de l'interopérabilité en interopérabilité technique, sémantique, organisationnelle et légale.

— le Level of Conceptual Interoperability Model (Tolk et al. (2007)) est la seule proposition à visée générale de notre sélection même s'il provient du domaine des systèmes de Modèle et de Simulation. Cette proposition décompose l'interopérabilité en interopérabilité technique, syntaxique, sémantique, pragmatique, dynamique, conceptuelle. L'interopérabilité est définie comme un modèle en couche, les différents types de mécanismes d'interopérabilité ne sont pas décrits mais des exemples d'applications du modèle à des problèmes en vue d'une implantation et d'une utilisation de ce cadre sont fournis.

— Noura et al. (2019) décrivent l'interopérabilité dans l'IoT ("Internet of Things") et décomposent l'interopérabilité en interopérabilité sémantique, syntaxique, des plateformes, des réseaux et des appareils. Les étapes d'implantation de l'interopérabilité ne sont pas définies mais une liste de différents mécanismes d'interopérabilité dans l'IoT est proposée (la standardisation, les adaptateurs / passerelles, la virtualisation, l'informatique nuagique (cloud et fog computing), etc...).

— Zeng (2019) décrit l'interopérabilité dans les systèmes d'organisation des connaissances ("Knowledge organisation systems") et décompose l'interopérabilité en interopérabilité système, syntaxique, structurelle et sémantique. Cette proposition ne propose pas d'étape d'implantation de l'interopérabilité mais décrit les mécanismes d'interopérabilité des standards de métadonnées et la problématique du mapping de modèles de métadonnées.

— Les principes FAIR ("Findable, Accessible, Interoperable, Reusable") énoncés par Wilkinson et al. (2016) définissent des principes généraux dans le but de mieux réutiliser des données et des connaissances. L'auteur ne décompose pas l'interopérabilité en types et la définit comme "la capacité des données ou des outils de ressources non coopératives à intégrer ou travailler ensemble avec un effort minimal". L'auteur définit trois principes liés à l'interopérabilité que les ressources de données, les outils, les vocabulaires et les infrastructures devraient présenter. Ces trois principes liés à l'interopérabilité sont :
  — I1 : Les (méta)données utilisent des langages formels, accessibles, partagés et largement applicables pour la représentation de savoir.
  — I2 : Les (méta)données utilisent des vocabulaires qui suivent les principes FAIR.

— I3 : Les (méta)données incluent des références qualifiées à d'autres (méta)données.
Ces principes sont décrits par l'auteur comme "précédant les choix d'implantation et ne suggèrent aucune technologie, standard ou solution d'implantation spécifique". Ces principes n'ont pas pour but de proposer une démarche d'implantation de l'interopérabilité mais uniquement à fournir aux éditeurs et aux gestionnaires de données une assistance pour l'évaluation de leurs choix d'implantation.

— Van Der Veer and Wiles (2008) décrivent l'interopérabilité dans les télécommunications et décomposent l'interopérabilité en interopérabilité technique, syntaxique, sémantique et organisationnelle. Cette proposition se base sur trois définitions différentes de l'interopérabilité et propose un processus pour répondre à l'interopérabilité dans les standards de l'ETSI ("European Telecommunications Standards Institute"). Cette approche se focalise uniquement sur la définition des standards et leur application dans un processus de standardisation

— Nilsson et al. (2008) décrivent les réponses à l'interopérabilité implantées dans le modèle de métadonnées standard Dublin Core. L'interopérabilité est décrite comme décomposable en un modèle en quatre couches. Les auteurs décomposent l'interopérabilité en interopérabilité sémantique formelle, syntaxique et des profils. Ces différentes couches sont décrites comme ne faisant pas l'objet d'un consensus. Les étapes de l'interopérabilité du modèle Dublin Core sont proposées avec celles pour réaliser l'interopération du modèle Dublin Core.

— Ambrosio and Widergren (2007) décrivent l'interopérabilité dans le contexte des entreprises et décomposent l'interopérabilité en interopérabilité technique, syntaxique, des réseaux, informationnelle, sémantique, organisationnelle et pragmatique. Cette proposition ne décrit pas les mécanismes d'interopérabilité et les étapes d'implantation.

— Zwegers (2003) décrit l'interopérabilité des entreprises mais vise une généricité. Cette proposition décompose l'interopérabilité en interopérabilité des communications, des données, des applications, des connaissances, des affaires et sémantique. Elle se base sur deux définitions de l'interopérabilité. Les étapes d'implantation ne sont pas décrites et l'interopérabilité est mise en place grâce à l'utilisation de standards communs.

— ATHENA (Berre et al. (2007)) décrit l'interopérabilité des processus inter-entreprises et décompose l'interopérabilité en interopérabilité des entreprises, des données, des informations, des processus, des services. Cette proposition se base sur la définition de l'IEEE ("Institute of Electrical and Electronics Engineers"), qui décrit l'interopérabilité comme "la capacité de deux ou plusieurs systèmes ou composants à échanger des informations et à utiliser les informations qui ont été échangées". Les étapes d'implantation sont décrites, la standardisation est vue comme un mécanisme d'implantation de l'interopérabilité et des exemples de solutions sont présentés.

— Kostoska et al. (2016) décrit l'interopérabilité du Cloud. Les auteurs distinguent l'interopérabilité du cloud, des sociétés, sémantique, technique, politique et humaine, des communautés, internationale, des plateformes de cloud et de la gestion de cloud. Cette proposition se base sur la définition de l'IEEE. Elle comporte un modèle conceptuel décrivant les aspects de l'interopérabilité, propose de mettre en œuvre l'interopérabilité en se focalisant sur les standards et fournit une méthode ainsi qu'un exemple d'implantation.

Un seul des travaux que nous avons sélectionnés (et que nous avons observé durant

notre exploration de la littérature) se positionne comme une proposition générique (Tolk et al. (2007)) et indépendante de tout contexte. Les dix autres se positionnent sur un domaine particulier. Plus précisément, les domaines couverts sont les suivants :

— Les services publics européens (U.E. dans le tableau) ;
— L'IoT ;
— Les systèmes d'organisation des connaissances (K.O dans le tableau) ;
— Les principes FAIR ;
— Les télécommunications (Télécom. dans le tableau) ;
— Les standards de métadonnées (Standard MD dans le tableau) ;
— Les entreprises ;
— Le Cloud.

Nous avons sélectionné trois travaux sur l'interopérabilité provenant d'un même domaine, le domaine des entreprises. Notre sélection permet d'observer s'il existe une variation interdomaines et intradomaines de la compréhension de l'interopérabilité.

Pour faciliter l'analyse des travaux existants, nous proposons le tableau 2.1. Chaque colonne correspond à un des travaux (cf. la légende du tableau) avec le domaine associé audessus. Pour chaque approche étudiée, nous avons observé la décomposition en types d'interopérabilité et indiqué par une croix les types d'interopérabilité qu'elle considère. Nous avons décomposé les types d'interopérabilité en deux groupes : les types d'interopérabilité définis selon les entités à rendre interopérables et les types d'interopérabilité définis selon l'objectif visé.

2.1.2 Les entités à rendre interopérables

Nous avons observé que les entités à rendre interopérables sont différentes selon le domaine ou la proposition sur l'interopérabilité. Les définitions de l'interopérabilité l'appliquent à trois catégories d'entités :

— Des systèmes, des composants ou des équipements pour l'IEEE, l'ISO ("International Organization for Standardization"), l'ETSI et le dictionnaire Merriam Webster's (cinq propositions utilisent cette définition (Noura et al. (2019); Zeng (2019); Van Der Veer and Wiles (2008); Zwegers (2003); Kostoska et al. (2016))) ;
— Des organisations pour l'EIF (une proposition utilise cette définition (Corcho et al. (2021))) ;
— Les données et les outils pour les principes FAIR (une proposition utilise cette définition (Wilkinson et al. (2016))).

En observant les entités décrites dans les différents travaux, nous notons un total de vingt-trois entités différentes. Il semble nécessaire de redéfinir les entités à rendre interopérables dans chacun de ces travaux.

Dans les différents types d'interopérabilité définis (cf. Tableau 2.1), les trois types d'entités présents dans les définitions de l'interopérabilité sont différents des entités utilisées pour définir les types d'interopérabilité.

Par exemple, les processus ne sont pas mentionnés dans les définitions tandis qu'une interopérabilité des processus est définie par Ambrosio and Widergren (2007). Nous comptons quinze entités qui ne sont pas représentées dans les définitions de l'interopérabilité (clouds, appareils, informations, réseaux, plateformes, processus, services, profils, savoirs, communications, politiques / humaine, plateformes de cloud, gestions des plateformes de cloud, communautés et applications). Les définitions proposées de l'interopérabilité ne représentent pas la variété d'entités à rendre interopérables.

Cette spécialisation des entités à rendre interopérables rend impossible l'adaptation à d'autres entités. Les solutions sur l'interopérabilité en considérant les organisations proposent des solutions et définissent des besoins pour les entreprises, avec des exemples de mécanismes qui leur sont applicables et qui ne sont pas applicables aux données et aux outils. Le changement d'entité implique un changement dans les besoins et les solutions. Une compréhension commune du concept d'interopérabilité doit définir de façon générique les entités à rendre interopérables mais aussi permettre une spécialisation de ces entités pour répondre à un besoin spécifique.

2.1.3 Les mécanismes d'interopérabilité

Aucune proposition ne définit les mécanismes permettant l'interopérabilité. Cependant, quatre propositions (Noura et al. (2019); Zeng (2019); Berre et al. (2007); Kostoska et al. (2016)) proposent clairement des exemples de mécanismes permettant la mise en place de l'interopérabilité. Ces propositions indiquent que l'utilisation de standards communs est une approche. Deux propositions (Noura et al. (2019); Zeng (2019)) proposent des exemples se basant sur la mise en place de passerelles entre les entités. En dehors des exemples expliqués, le manque de définition claire d'un mécanisme d'interopérabilité et des approches existantes pour ces mécanismes ne permet pas de catégoriser les outils qui ne font pas partie des exemples proposés par ces auteurs. Avec ces seuls travaux, il n'est pas possible de comprendre les mécanismes d'interopérabilité d'un nouveau contexte.

2.1.4 Les démarches d'implantation

Deux propositions décrivent la démarche d'implantation de l'interopérabilité. L'une est un modèle générique (Tolk et al. (2007)) qui ne propose pas la méthode d'application de cette démarche et ne propose pas de définition des mécanismes pour réaliser cette implantation. L'autre vise à répondre à l'interopérabilité du modèle Dublin Core (Nilsson et al. (2008)) ne permettant pas une adaptation de cette démarche d'application à d'autres problèmes d'interopérabilité. Les autres propositions ne proposent pas d'implantation de l'interopérabilité. Aucun cadre ne fournit de compréhension de l'implantation de l'interopérabilité et son application en même temps. L'implantation de l'interopérabilité n'est pas possible en l'état.

2.1.5 Les types d'interopérabilité supportés

Nous observons que les différents types d'interopérabilité définis dans l'état de l'art peuvent se décomposer en deux catégories (cf Tableau 2.1) : les types d'interopérabilité définis selon les entités à rendre interopérables (par exemple, l'interopérabilité des entreprises) et les types d'interopérabilité définis par l'objectif de cette interopérabilité (par exemple, l'interopérabilité légale selon Corcho et al. (2021) est "la capacité à combiner des données de plusieurs sources sans conflits entre les licences sur les données").

Six travaux sur l'interopérabilité intègrent des types d'interopérabilité définis selon le type d'entités, avec un total de dix-huit types d'interopérabilité selon les entités à rendre interopérables. Nous observons que toutes les solutions du domaine des entreprises intègrent plusieurs types d'interopérabilité selon les entités à rendre interopérables. Les domaines de l'IoT, du cloud, des entreprises et le standard Dublin Core portent une grande attention aux différentes entités. Une compréhension commune **générique** doit permettre

de proposer une compréhension pour l'ensemble des entités à rendre interopérables afin de répondre à l'ensemble des problèmes d'interopérabilité.

Huit travaux sur l'interopérabilité sur les onze intègrent des types d'interopérabilité selon les objectifs, avec un total de treize types d'interopérabilité définis selon les objectifs. Les domaines des services publics européens, génériques, de l'organisation des connaissances et des télécommunications portent une grande attention aux objectifs de l'interopérabilité. Une compréhension commune **générique** doit permettre de proposer une solution à l'ensemble des problèmes d'interopérabilité qui peuvent être rencontrés, afin de répondre à l'ensemble des problèmes d'interopérabilité.

Ces différents travaux sur l'interopérabilité ne décomposent pas l'interopérabilité avec les mêmes types d'interopérabilité. Au cœur du domaine des entreprises, les entités à rendre interopérables sont différentes (cf. les propositions 8, 9 et 10 dans le tableau) malgré la proximité des problèmes soulevés. Ces travaux sont spécialisés sur un problème spécifique. Ces travaux ne sont pas généralisables à d'autres problèmes à cause de cette spécialisation. Cette spécialisation des solutions sur l'interopérabilité explique un **manque d'exhaustivité** de prise en compte des problématiques de l'interopérabilité observés avant (Zwegers (2003); Rezaei et al. (2014a)).

Neuf travaux sur les onze intègrent dans leur décomposition un type d'interopérabilité répondant à un problème sémantique (interopérabilité sémantique, sémantique formelle ou des informations). Neuf travaux sur les onze intègrent dans leur décomposition un type d'interopérabilité répondant à un problème technique (interopérabilité technique, syntaxique ou système). Les problèmes techniques et sémantiques sont largement partagés dans les travaux sur l'interopérabilité. Une compréhension commune **générique** doit décomposer l'interopérabilité en deux composantes : une composante technique et une composante sémantique.

2.1.6 Bilan

Nous avons observé qu'**il n'existe pas de compréhension commune du concept d'interopérabilité**. Cette compréhension commune est décrite comme nécessaire par les chercheurs (Rezaei et al. (2014a)) et son absence provoque une non-exhaustivité des réponses apportées à la problématique d'interopérabilité (Zwegers (2003); Rezaei et al. (2014a)).

Notre analyse comparative nous a permis d'observer plusieurs points :

— L'interopérabilité est un concept complexe, au vu des différents types d'interopérabilité définis au sein des travaux sur l'interopérabilité.
— L'objectif et les entités à rendre interopérables sont des facteurs majeurs dans la définition d'un problème d'interopérabilité (cf. Section 2.1.5).
— Les propositions sur l'interopérabilité ne sont pas suffisamment génériques pour être appliquées à d'autres contextes que ceux prévus par ces propositions (cf. Section 2.1.2).
— Les mécanismes permettant l'interopérabilité ne sont ni définis, ni caractérisés. Des exemples de mécanismes sont proposés sans définir pour autant pourquoi ils permettent l'interopérabilité ni leur caractérisation (cf. Section 2.1.3).
— Les propositions actuelles ne proposent que rarement une démarche d'implantation de l'interopérabilité. Quand cette démarche est définie, la méthode d'application n'est pas proposée (cf. Section 2.1.4).

Cette analyse comparative des travaux sur l'interopérabilité nous a permis d'observer

qu'une compréhension commune doit :

— rendre compte de la complexité de l'interopérabilité, par exemple à travers un modèle en couches et en illustrant les deux composantes de l'interopérabilité (sémantique et technique) ;
— être **exhaustive** : en définissant l'interopérabilité, nous attendons que soient précisés les types d'entités qui peuvent être considérées, la démarche d'implantation, les mécanismes permettant l'interopérabilité et leur caractérisation ;
— être **générique**, pour pouvoir être appliquée à l'ensemble des problèmes d'interopérabilité définis, en considérant les différentes entités et la diversité des objectifs possibles.

Dans la section suivante, nous explorons les différentes solutions sur les données de recherche dans la Science Ouverte afin de caractériser les entités que nous souhaitons rendre interopérables.

2.2 Les plateformes de données ouvertes de la recherche

La Science Ouverte se définit comme "un accès ouvert aux connaissances partagées et développées à travers un réseau de collaboration dans le monde de la recherche scientifique" (Vicente-Saez and Martinez-Fuentes (2018)). Ce réseau de collaboration se réalise à travers le partage de données entre les différents producteurs et utilisateurs de données de recherche. Nous définissons le terme générique Plateforme de Données de Recherche Ouverte (PDRO) pour désigner l'ensemble des solutions nécessaires à la gestion, l'accès ou la recherche de données de recherche dans la Science Ouverte. Pour permettre l'exploration de données de recherche, il est nécessaire de mettre en place une interopérabilité entre ces PDRO. Les PDRO sont décrites comme nombreuses et variées. La connaissance des entités à rendre interopérables est essentielle comme vu dans la section 2.1.

L'objectif de cette section est de caractériser ces PDRO pour comprendre les verrous auxquels se heurte l'implémentation de l'interopérabilité des PDRO. Pour réaliser cette analyse, nous suivons les deux composantes de l'interopérabilité, technique et sémantique.

Dans la section 2.2.1, nous proposons une classification des PDRO en deux catégories : (1) les portails de données et (2) les architectures de gestion de données. Nous présentons les portails de données sélectionnés, en les distinguant selon les types de données auxquels ils permettent l'accès. Ensuite, nous présentons les architectures de gestion de données. En section 2.2.2, nous réalisons une analyse comparative selon la composante technique, avec les API de communication et les technologies utilisées. Nous réalisons ensuite une analyse comparative selon la composante sémantique, en étudiant les modèles de métadonnées utilisés et les formats de ces métadonnées.

2.2.1 Présentation des plateformes de données de recherche ouvertes

Parmi les solutions de gestion et d'accès aux données de recherche, nous observons les portails de données, les plateformes de données et les dépôts de données de recherche ("data research repository").

Le portail de données possède plusieurs définitions. Un portail de données peut être "une collection de métadonnées, combinée à un système de gestion et de recherche de

données, qui aide les analystes et les autres utilisateurs de données à trouver les données dont ils ont besoin, servant comme un inventaire des données disponibles, et fournissant des informations pour évaluer l'adaptation de ces données à une utilisation spécifique" [1]. Mais le portail de données peut aussi être défini comme un "front office" d'une plateforme de gestion de données (Dymytrova and Paquienséguy (2017)). Selon Dymytrova and Paquienséguy (2017), la plateforme associée à un portail a pour objectif "l'hébergement, la gestion des données et l'interopérabilité des données". Ces deux définitions s'accordent sur deux fonctionnalités présentes sur ces portails : l'accès aux données et la recherche de données. En revanche, le stockage des métadonnées n'est pas une composante fonctionnelle commune. De plus, les informations sur les plateformes sur lesquelles se basent les portails de données sont peu souvent explicitées. Il est donc difficile de les catégoriser correctement.

Le second type de solution est la plateforme de données. Une plateforme de données est définie comme une "solution complète de bout en bout pour toutes [les] données. Une [..] plateforme de données peut ingérer, traiter, analyser et présenter les données générées par tous les systèmes et infrastructures au sein de votre organisation." [2]

Le dernier type est le dépôt de données de recherche ("data research repository"). Les dépôts de données de recherche sont définis par Uzwyshyn (2016) comme "de grandes infrastructures de bases de données mises en place pour gérer, partager, accéder et archiver les données de recherche". Cette plateforme intègre les fonctionnalités des portails de données et des plateformes de gestion de données. Elle semble être un entre-deux entre les deux premières solutions, ou une combinaison de celles-ci. Il nous paraît difficile de catégoriser correctement les dépôts de données de recherche dans l'une ou l'autre de ces catégories des deux premières solutions. Pour permettre la catégorisation des solutions avec les informations qui sont disponibles, nous distinguons **fonctionnellement** ces PDRO en deux types de solutions :

— **Les portails de données**, que nous définissons comme une solution permettant l'accès aux données de recherche ainsi que la recherche de données de recherche avec laquelle les utilisateurs interagissent ;

— **Les architectures de gestion de données**, que nous définissons comme une solution permettant le stockage, l'archivage, le traitement et l'analyse des données de recherche ainsi que des métadonnées.

La sélection des portails et des architectures de gestion de données que nous avons réalisée est reportée dans le tableau 2.2. Les cases vides sont dues à un manque d'informations sur ces PDRO. Pour la composante technique, nous avons pris en compte la présence ou l'absence d'API de communication permettant l'accès aux services d'une PDRO et les technologies utilisées nécessaires à la mise en place des services d'accès ou de gestion de données. Pour la composante sémantique, nous avons cherché les modèles de métadonnées utilisés, soit pour la recherche de données dans les portails, soit pour la modélisation des métadonnées dans les architectures de gestion de données, et les formats de métadonnées, permettant la structuration des métadonnées quand elles transitent.

1. www.alation.com/blog/what-is-a-data-catalog/
2. www.splunk.com/en_us/blog/learn/data-platform.html

2.2.1.1 Les portails de données

OpenDataMonitor [3] et Dataportal.org [4] référencent près de 800 portails de données ouverts, ce qui atteste de la grande quantité de portails existants. Les portails de données fournissent généralement aux utilisateurs une interface graphique intégrant des outils de recherche et d'accès aux données de recherche. Ces portails permettent l'accès à un catalogue de données décrivant l'ensemble des jeux de données de recherche conservés sur une architecture de gestion de données. Nous catégorisons ces portails en fonction du type de données ouvertes qu'ils gèrent : les données ouvertes, les publications scientifiques ou les données gouvernementales.

Les portails de données dédiés aux données ouvertes Les portails privés, comme OpenDataSoft [5] ou Figshare [6], ou les portails publics généralement déployés par et pour des universités ou des équipes de recherche, comme dat@UBFC [7], fournissent une interface graphique utilisateur avec un système de recherche d'information riche. Ces portails peuvent être liés à une discipline ou une communauté, comme FigShare proposant une gestion des données des Sciences Sociales et Humanités, ou être généraux comme le portail OpenDataSoft.

Les portails privés se basent généralement sur des solutions de gestion de données qu'ils développent eux-mêmes comme OpenDataSoft, ou ne permettent pas d'avoir d'information sur les technologies et / ou les modèles utilisés. Les portails publics ne fournissent que peu souvent des informations concernant les solutions sur lesquelles ils se basent.

Nous avons sélectionné la plateforme ouverte d'OpenDataSoft pour illustrer cette catégorie, faute d'informations sur les autres plateformes.

Les portails de données dédiés aux publications scientifiques Cette catégorie de plateformes est la moins répandue.

En effet, la grande notoriété de certains portails de données sur les publications scientifiques n'incite pas au développement de nouveaux portails de ce type. Google Scholar [8] est le meilleur candidat pour représenter cette catégorie de solutions de gestion de données de la Science Ouverte. Google Scholar est déjà largement répandue dans la recherche scientifique.

Mais l'architecture de Google Scholar n'est pas ouverte. Les différents processus mis en place derrière l'interface graphique du portail ne sont pas accessibles. Aucune API de communication n'est proposée. Cette absence d'API empêche l'automatisation des communications, par exemple pour faire de la collecte de métadonnées, ou l'utilisation d'autres outils que ceux proposés par Google Scholar. Parmi les autres portails privés de publications les plus connus et utilisés, nous trouvons Web of Science [9]. Parmi les portails publics, citons PubMed [10] géré par le gouvernement américain pour l'accès à la littérature scientifique biomédicale.

3. opendatamonitor.eu
4. dataportals.org
5. public.opendatasoft.com/
6. figshare.com/
7. search-data.ubfc.fr/
8. scholar.google.com/
9. clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/
10. pubmed.ncbi.nlm.nih.gov/

HAL [11] est une solution complexe au service de la communauté française. HAL comprend un portail de saisie et de recherche de publications scientifiques et une architecture de gestion de données. Les modèles utilisés dans HAL ne sont pas clairement indiqués et la définition de métadonnées passe par des interfaces graphiques proposant des formulaires simples pour les utilisateurs [12]. Cependant, la documentation [13] indique que des modèles de métadonnées standards sont ré-utilisés par HAL au sein d'un modèle ad-hoc. La documentation vers le référentiel de métadonnées n'est pas accessible (au 7 août 2024 [14]). Les archives ouvertes intégrées au sein de HAL doivent répondre à une liste d'exigences, dont les exigences OpenAIR pour les métadonnées [15].

Nous avons sélectionné la plateforme HAL et le portail de la solution ISTEX (cf. Section 2.2.1.2 pour illustrer les portails de données des publications scientifiques.

Les portails de données dédiés aux données gouvernementales Les portails de données gouvernementales sont très nombreux et sont déployés par une grande variété d'institutions publiques.

Ils permettent d'accéder aux informations d'institutions publiques qui peuvent être utilisées dans certaines disciplines, comme les Sciences Sociales avec les statistiques de l'INSEE. Nous retrouvons ce type de portail déployé par tous les pays du monde mais aussi par des institutions supranationales comme l'Union Européenne. Par exemple, l'EOSC (European Open Science Cloud) est un projet européen dont l'objectif est d'interconnecter des sources de données afin d'en fournir un accès transparent. L'EOSC Marketplace était le portail de données de ce projet, mais le projet étant arrêté (au 7 Août 2024 [16]), nous ne pouvons pas inclure cette solution dans notre analyse. L'EOSC Resource Catalog est le composant permettant la gestion du catalogue de données de l'EOSC [17]. Des critères sont définis pour s'intégrer au cadre de l'EOSC et un ensemble de modèles de métadonnées sont définis pour servir de modèle pivot afin d'adapter les modélisations de métadonnées des différentes sources [18]. Ce projet vise à interconnecter des plateformes existantes et non fournir une architecture technique pour la gestion de données ou les portails de données [19].

Le gouvernement français propose un portail de données avec data.gouv.fr [20], qui intègre plusieurs catalogues, notamment recherche.data.gouv.fr pour les données de recherche. Les municipalités proposent aussi des portails pour l'accès aux données en lien avec leur ville, comme le portail de la ville de Paris [21]. Enfin, les services publics fournissent aussi des portails de données, comme le portail de la SNCF [22] qui donne accès aux données de mobilité.

Différentes plateformes sont aussi développées dans d'autres pays, comme les plate-

11. about.hal.science/
12. doc.hal.science/en/x2hal/verifier-et-completer-les-metadonnees/
13. about.hal.science/principles/
14. api.archives-ouvertes.fr/ref/metadata
15. www.coalition-s.org/plan-s-practical-advice/#Requirements_for_Open_Access_Repositories
16. eosc-portal.eu/
17. eoscfuture.eu/eosc-resource-catalogue
18. openaire-guidelines-for-literature-repository-managers.readthedocs.io
19. eoscfuture.eu/wp-content/uploads/2022/04/EOSC-Core.pdf
20. data.gouv.fr
21. opendata.paris.fr/
22. ressources.data.sncf.com

formes du gouvernement américain [23] ou du gouvernement britannique [24]. L'ouverture des données par les autres niveaux institutionnels n'est pas une spécificité de la France, avec notamment un portail de la ville de New-York [25]. Nous observons aussi des plateformes proposées par l'Union Européenne comme data.europa [26]. Cette plateforme étend la gestion de données aux jeux de données et aux publications scientifiques.

Pour notre étude, nous avons sélectionné plusieurs portails de données gouvernementales à plusieurs niveaux :

— un portail déployé par des institutions supranationales, avec l'EOSC Resource Catalogue, seule entité encore en fonctionnement du portail de données EOSC Marketplace ;
— un portail déployé par le gouvernement, provenant de plusieurs pays (France, États-Unis et Royaume-Uni) ;
— un portail déployé par les communes, provenant de plusieurs pays (Paris et New-York) ;
— un portail déployé par des services publics, avec le portail de la SNCF.

2.2.1.2 Les architectures de gestion de données

Les architectures de gestion de données visent à stocker, archiver, traiter et analyser les données. Même si ce sont les portails de données qui s'en servent majoritairement, ces architectures intègrent un système de gestion de métadonnées, à la fois pour conserver le catalogue de données, mais aussi permettre le fonctionnement interne à ces architectures.

Les types d'architectures sont variés.

Vuong et al. (2018) proposent une architecture de gestion de données centrée sur des bases de données relationnelles ouvertes via une interface graphique web et une API RESTful. Cette solution intègre un modèle relationnel créé sur mesure pour les besoins applicatifs. Elle vise à trouver les coopérations entre chercheurs Vietnamiens en sciences sociales.

Castro et al. (2022) proposent une architecture de référence pour une gestion FAIR du Big Data en vue de l'analyse d'un grand volume de données de la Science Ouverte. Cette solution se base sur des entrepôts de données pour la gestion des métadonnées, appelés "metadata warehouse". Pour la gestion des données, cette solution utilise une architecture de lac de données basée sur la technologie Hadoop.

Beheshti et al. (2018) définissent une architecture de "Knowledge Lake", que nous traduisons en lac de connaissances. Cette architecture dérivée du lac de données est basée sur une intégration de bases de données NoSQL ainsi que des bases de données relationnelles pour la gestion des connaissances et des métadonnées, en vue de produire un graphe de connaissances. Les métadonnées permettent de suivre le cycle de vie des données dans cette architecture et enregistrent les différentes opérations effectuées sur ces données.

Killough (2018) propose une solution de Business Intelligence adaptée à la Science Ouverte avec un cube de données à travers l'initiative Open Data Cube. Cette architecture est une plateforme d'analyse et de gestion de données ouvertes pour la gestion de données satellites. Les métadonnées sont formalisées grâce à un modèle de métadonnées sur mesure, flexible, ne suivant pas de standard existant.

23. `data.gov`
24. data.gov.uk
25. `opendata.cityofnewyork.us/`
26. `data.europa.eu/`

Ermilov et al. (2015) proposent de gérer des données de la Science Ouverte à l'aide d'une base de données graphe ou triplestore. L'objectif est de transformer les métadonnées disponibles dans des bases de données relationnelles internes d'universités en graphes RDF, notamment grâce à l'utilisation d'un système de mapping. Cette solution utilise l'ontologie VIVO [27] qui a été enrichie en VIVO+ afin de compenser les lacunes de VIVO pour l'adapter aux métadonnées des jeux de données des universités. L'ontologie VIVO permet de représenter les chercheurs dans le contexte de leur expérience, de leurs résultats, de leurs intérêts, de leurs réalisations et des institutions qui leur sont associées.

Dooley et al. (2018) décrivent la plateforme AGAVE. Cette plateforme est une solution pour la gestion de données dans le domaine de la biologie des plantes. Elle permet de rendre réutilisables ces données et intègre de nombreuses applications ainsi que des clusters de Calcul Haute Performance ("High-Performance Computing" - HPC). La gestion des données est réalisée grâce à une base de données MongoDB avec, *a priori*, un modèle de métadonnées non standardisé créé sur mesure pour les besoins applicatifs.

Aryani et al. (2020) définissent une architecture théorique dans le but de rendre interopérables les graphes de connaissances de la Science Ouverte. L'interopérabilité de ces graphes est décrite comme nécessaire et les auteurs indiquent qu'un "framework" de compréhension de l'interopérabilité appliqué aux graphes de connaissances doit être développé. Les auteurs proposent une architecture fonctionnelle de haut niveau pour décrire la structuration des solutions pour gérer ces graphes de connaissances qui sont basés sur les bases de données graphes.

L'infrastructure de HAL [28] intègre une grande variété de solutions de gestion de données (MySql, Apache Solr, Virtuoso, NAS) mais ne fournissant pas d'API ouverte aux utilisateurs. L'accès aux données de HAL est réalisé via le portail de HAL, comme le seul point d'accès défini dans la documentation.

ISTEX [29] est un projet complexe comprenant un portail de données pour accéder à une grande collection de publications scientifiques et une architecture de gestion de données permettant la mise en place de fouilles de données, décrites comme la fonctionnalité cœur de cette solution. Cette solution se base sur la technologie Elasticsearch. La gestion des métadonnées est basée sur l'utilisation d'un graphe des données dont le modèle ad-hoc peut être aligné avec des modèles externes [30]. Dans l'analyse de cette solution, nous avons distingué l'architecture et le portail proposé par l'ISTEX (cf Tableau 2.2).

Harvard Dataverse est une architecture de données et un portail de données permettant l'accès à de nombreuses données de recherche dans divers domaines. Cette solution est basée sur la technologie Dataverse et utilise une modélisation des métadonnées ad-hoc réalisée pour les besoins applicatifs [31]. Cette modélisation est décomposée en blocs de métadonnées, annoncés comme compatibles avec d'autres modèles de métadonnées comme DDI, DataCite ou Dublin Core. Harvard Dataverse propose une API REST pour interagir avec l'architecture et une interface graphique Web pour les utilisateurs du portail.

Pour conclure, les architectures de gestion de données de la Science Ouverte se basent sur des architectures et sur des paradigmes de gestion de données différents. Leurs spécificités impactent les processus de gestion, les données pouvant être gérées ou les fonctionnalités proposées. Nous avons sélectionné chacune des architectures de gestion de données décrites

27. github.com/vivo-ontologies/vivo-ontology
28. about.hal.science/infrastructure/
29. www.istex.fr/
30. data.istex.fr/
31. guides.dataverse.org/en/latest/user/appendix.html

pour notre analyse.

2.2.1.3 Conclusion

Nous avons exploré vingt-cinq PDRO. Nous avons conservé vingt de ces PDRO, par manque d'informations sur les cinq autres. Nous avons observé que les informations décrivant les technologies ou les modèles de métadonnées utilisés dans ces PDRO sont difficiles à trouver ou peuvent ne pas être fournies. Ce constat est partagé dans des revues de la littérature comme (Dymytrova and Paquienséguy (2017)). Notre sélection inclut des portails de données et des architectures de gestion de données utilisant une grande variété de technologies. Nous proposons dans la suite une analyse comparative de ces PDRO.

2.2.2 Analyse comparative des PDRO

Nous proposons de comparer les PDRO selon les deux composantes de l'interopérabilité :

— l'interopérabilité technique : nous comparons les solutions techniques des PDRO. Les API de communication permettent la communication avec la plateforme et les technologies utilisées permettent de fournir les différents services, comme la définition d'une interface graphique Web ou la mise en place d'une fouille de données.

— l'interopérabilité sémantique : nous comparons les solutions permettant la description des données de recherche. Les modèles de métadonnées permettent de définir les informations et les formats de métadonnées permettent le transport de ces métadonnées.

Notre sélection vise à couvrir exhaustivement les types de plateformes de gestion de données que nous avons trouvées. Nous avons sélectionné au moins un portail gérant chaque type de données, et des portails proposés à différents niveaux (public / privé, supranational, national ou communal).

2.2.2.1 La composante technique - les API de communication

Notre première observation concerne la variété des API de communication. Sur les vingt plateformes de données de recherche ouvertes, neuf types d'API différents sont utilisés : REST, Apache Solr, Sword, OAI-PMH, Kafka, une bibliothèque python, SPARQL, APIM et SOAP. L'API REST est celle qui est la plus représentée et la plus utilisée, avec treize plateformes proposant une API REST.

Cette diversité d'API empêche les différentes plateformes de communiquer. Les API basées sur des langages de requêtage de système de gestion de base de données sont peu représentées, avec seulement deux plateformes (SPARQL). L'API REST est utilisée par l'ensemble des portails de données, montrant une compatibilité avec la mise à disposition d'accès aux ressources d'une plateforme. La technologie des API REST permet un développement flexible et l'utilisation de différents protocoles de communication et outils de requêtage.

Ensuite, nous observons qu'en utilisant la même technologie d'API basée sur REST, les implémentations de ces API peuvent varier. Nous observons neuf implémentations d'API différentes : uData, OpenDataSoft API, CKAN, ISTEX et les API REST sur mesure. Chaque implémentation gère de façon particulière les URLs, le type de requête utilisée pour chaque chemin et par conséquent les capacités de communication. La di-

Proposition	API	Modèle de métadonnées	Technologie(s)	Format de métadonnées
Portail de données				
data.gouv.fr	REST (uData)	DCAT, CKAN, DKAN	uData (Moissonnage)	JSON
SNCF	REST (OpenDataSoft API)	Ad-Hoc	OpenDataSoft	JSON
Opendata Paris	REST (OpenDataSoft API)	Ad-Hoc	OpenDataSoft	JSON
data.gov (Etats-unis)	REST (CKAN)	US-DCAT, ISO 19115,	API Umbrella / CKAN	JSON, XML
EOSC - Resource Catalogue	API REST, Apache Solr, Sword, OAI-PMH	DataCite, OpenAire	OAI-PMH	
HAL - portail		Dublin Core, RDF, FOAF, SKOS, BILBO, Fabio, ad-hoc		
Istex - portail	API ISTEX (REST), SparQL	Dublin Core		XML, MODS
Open Data Platform	REST (CKAN)	Ad-Hoc	OpenDataSoft	
New-York	REST	Ad-Hoc	Socrata APIs	
data.gov.uk (Royaume-Unis)	REST (CKAN)	ISO 19139, GEMINI, DCAT	CKAN	XML
OpenDataSoft	REST (OpenDataSoft API)	Ad-Hoc		JSON
Architecture de gestion de données				
HAL - architecture			Apache Solr, MySQL, Virtuso, NAS	
ISTEX - architecture		Ad-Hoc	Elasticsearch	
Vuong et al. (2018)	REST	Ad-Hoc	BD Relationnelle	
Castro et al. (2022)	Kafka	Ad-Hoc	Lac de données, entrepôt de données	
Beheshti et al. (2018)	REST	Ad-Hoc	Lac de données BD NoSQL et relationnelle	
Killough (2018)	Librairie Python	Ad-Hoc	Cube de données	YAML
Ermilov et al. (2015)	SPARQL	VIVO+	Triplestore	
Aryani et al. (2020)			Base de données graphe	
Dooley et al. (2018)	REST, APIM, SOAP	Ad-Hoc	Base de données orientée document, cluster HPC	

TABLE 2.2 – Comparatif des plateformes de données de recherche ouvertes

versité des API, et donc de gestion des communications, est donc un frein important à l'interopérabilité de PDRO.

2.2.2.2 La composante technique - les technologies

L'analyse des technologies porte sur des caractéristiques différentes selon les deux types de plateformes que nous étudions, car elles n'ont pas les mêmes objectifs fonctionnels.

Les technologies et les solutions techniques utilisées pour déployer des portails de données sont généralement peu documentées et les informations sur ces solutions sont difficiles à trouver (constat partagé par Dymytrova and Paquienséguy (2017)). Nous identifions les six technologies suivantes : uData, OpenDataSoft, API Umbrella, CKAN, OAI-PMH, Socrata API. Cependant, il est possible de s'abstraire des différences entre ces technologies grâce à la mise en place d'une API REST, ce qui est le cas dans tous les portails. En effet, une API REST permet une utilisation transparente des technologies sous-jacentes.

Les technologies utilisées par les architectures de gestion de données sont présentées de manière plus détaillée par les auteurs. En effet, ces technologies sont le cœur des propositions afin d'obtenir les fonctionnalités voulues. Le panel de solutions technologiques est vaste. Nous trouvons :

— des systèmes de gestion de base de données relationnelles, NoSQL et triplestore ;
— des architectures Big Data, avec les lacs de données, les entrepôts de données, les cubes de données et les clusters HPC ;
— une solution réseau avec la technologie NAS (Network Attached Storage).

Nous observons que quatre architectures de gestion de données sont des architectures complexes intégrant plusieurs de ces technologies en même temps. Ces architectures traitent la gestion de données de manière différente, ce qui conditionne les types de données gérés et normalement les mécanismes d'accès à ces solutions. À l'instar des architectures techniques des portails de données, les API permettant l'interaction d'utilisateurs sont peu documentées.

Les accès à ces différentes technologies sont réalisés via les API. Nous observons que les API REST permettent l'accès à un grand nombre de technologies différentes. Du point de vue de l'interopérabilité, le problème de la variété de technologies n'impacte donc pas l'interopérabilité des PDRO grâce à ces API de communication qui assurent l'accès à de nombreuses technologies.

2.2.2.3 La composante sémantique - les modèles de métadonnées

Il existe un flou sur la distinction entre modèle de métadonnées et vocabulaire. Par exemple, les vocabulaires RDF, FOAF, SKOS, BILBO et Fabio sont des vocabulaires définis dans la documentation de HAL sur les modèles de métadonnées [32]. Pour cette analyse, nous les considérons comme des modèles de métadonnées.

Ensuite, sur les vingt plateformes comparées, douze utilisent un modèle de métadonnées spécifiquement défini pour les besoins applicatifs (cf. "modèle ad-hoc" dans le tableau 2.2). Chaque modèle ad-hoc de notre sélection est différent des autres. Ainsi, cela représente un total de vingt-cinq modèles de métadonnées différents pour vingt PDRO au total.

Nous observons que les modèles de métadonnées de portails de données sont en grande partie standardisés. Sur onze portails de données, quatorze modèles de métadonnées stan-

32. about.hal.science/principles/

dardisés sont réutilisés : DCAT, CKAN, DKAN, US-DCAT, ISO 19115, Dublin Core, RDF, FOAF, SKOS, BILBO, Fabio, GEMINI.

Ainsi, sur ces vingt plateformes, trente modèles de métadonnées sont utilisés. Nous notons que l'utilisation de modèles standards ne permet pas de réduire la variété de modèles et donc d'améliorer l'interopérabilité. Pour un même modèle standard, il peut exister plusieurs versions de ce même modèle (cf. DCAT et US-DCAT). La standardisation ne permet pas de répondre à la problématique d'interopérabilité pour les portails de données à l'échelle de la Science Ouverte et une autre approche doit être prise pour mieux rendre interopérables ces plateformes. La variété de modèles de métadonnées empêche une compréhension commune des informations sur les données de recherche et la mise en place d'un accès unifié aux métadonnées. Or, l'interopérabilité requiert de répondre à l'ensemble de ces besoins et doit être possible malgré cette très grande variété de modèles de métadonnées.

2.2.2.4 La composante sémantique - les formats d'échange de métadonnées

Les formats de métadonnées permettent de définir la structure des messages pour le déplacement de ces métadonnées lors d'une recherche de données utilisant ces métadonnées. Ces formats sont généralement utilisés par les API de communication pour formater les réponses aux recherches de données. Or peu de plateformes renseignent les formats utilisés. Les quatre formats différents que nous avons trouvés (JSON, YAML, XML, MODS) sont cependant compatibles : il est simple de transformer des documents d'un format dans l'autre sans contrainte particulière. Les outils de transformation sont nombreux et leur utilisation peu coûteuse. La variété particulière des formats de métadonnées ne pose donc pas de problème pour mettre en place l'interopérabilité.

2.2.2.5 Bilan

La Science Ouverte se définit comme "un accès ouvert aux connaissances partagées et développées à travers un réseau de collaboration dans le monde de la recherche scientifique" (Vicente-Saez and Martinez-Fuentes (2018)). Ce réseau de collaboration se réalise à travers un partage de données entre producteurs et utilisateurs de données via des plateformes et des portails. Nous définissons le terme générique *Plateformes de Données de Recherche Ouvertes* (PDRO) pour désigner une solution visant soit la gestion, soit l'accès soit le partage des données de recherche dans la Science Ouverte. Pour permettre une recherche unifiée de données de recherche, il est nécessaire de mettre en place une interopérabilité entre ces PDRO. Les PDRO sont nombreuses et variées. La connaissance des entités est essentielle pour les rendre interopérables, comme vu dans la section 2.1. Pour comprendre les problématiques d'interopérabilité de ces PDRO, nous avons réalisé une analyse comparative de ces différentes PDRO. Nous catégorisons des PDRO en portails de données - permettant l'accès et la recherche de données - et en architectures de gestion de données - permettant le stockage, l'archivage, l'analyse et le traitement des données et des métadonnées. Nous avons sélectionné différentes plateformes qui soient représentatives de la variété des PDRO.

L'analyse comparée de ces plateformes nous a permis d'extraire quatre conclusions :

— Les API de communication possèdent une grande variété qui empêche la communication entre les PDRO.

— Les technologies n'ont pas d'impact sur l'interopérabilité, grâce à la capacité d'abstraction des API offrant une utilisation transparente de ces outils pour les utilisa-

teurs.
— Les modèles de métadonnées sont très variés. La standardisation de ces modèles ne réduit pas leur nombre et leur variété, ce qui est un problème majeur d'interopérabilité.
— Les formats d'échange de métadonnées sont peu nombreux et ne posent pas de problème d'interopérabilité.

Une solution d'interopérabilité doit donc répondre à deux variétés observées dans les plateformes de gestion de données : (1) la variété des API de communication et (2) la variété des modèles de métadonnées.

2.3 Conclusion

L'interdisciplinarité dans la Science Ouverte est perçue comme nécessaire à la fois pour enrichir les processus de création de connaissances par le développement de réponses innovantes à des questions de recherche et pour certaines communautés qui sont intrinsèquement interdisciplinaires. Pour permettre le déploiement à large échelle de cette interdisciplinarité, l'exploration de données de recherche doit être rendue accessible à l'ensemble des acteurs de la Science Ouverte en réalisant un partage d'informations sur ces données. Nous définissons le terme *Plateforme de Données de Recherche Ouverte* (PDRO) pour désigner l'ensemble des solutions visant la recherche, l'accès, la gestion et/ou l'analyse des données de recherche.

Pour comprendre les enjeux de ce partage d'informations, nous avons étudié le concept d'interopérabilité. Nous n'avons trouvé aucune solution expliquant l'interopérabilité des PDRO. De nombreux travaux sur l'interopérabilité ont été proposés mais il n'existe pas de compréhension commune, alors qu'elle est jugée nécessaire par les chercheurs (Rezaei et al. (2014a)). L'analyse de plusieurs approches de l'interopérabilité nous a permis de retenir les points suivants :
— L'interopérabilité est un concept complexe ayant deux composantes : la composante technique et la composante sémantique.
— L'objectif et les entités à rendre interopérables sont des facteurs majeurs dans la définition d'un problème d'interopérabilité (cf. Section 2.1.5).
— Les propositions sur l'interopérabilité ne sont pas applicables au problème d'interopérabilité des PDRO par manque de généricité (cf. Section 2.1.2).
— Ces travaux ne définissent ni les mécanismes permettant l'implémentation de l'interopérabilité ni leur caractérisation (cf. Section 2.1.3).
— Ces travaux ne définissent pas la démarche d'implémentation de l'interopérabilité ni la méthode d'application de cette démarche (cf. Section 2.1.4).

Dans le but de comprendre les besoins de l'interopérabilité des PDRO, nous avons analysé les PDRO suivant les deux composantes de l'interopérabilité : (1) la composante technique avec les API de communication et les technologies de ces PDRO et (2) la composante sémantique avec les modèles de métadonnées et les formats de ces métadonnées. Cette analyse nous a permis d'extraire plusieurs points :
— La variété de modèle de métadonnées et la variété des API de communication sont des freins majeurs à l'implantation de l'interopérabilité ;
— Les API offrent des capacités d'accès unifié à une large variété de technologies et la variété des formats d'échange de métadonnées est faible. Ces deux variétés ne sont pas des freins à l'implantation de l'interopérabilité.

Pour répondre à la problématique de l'interopérabilité des PDRO, nous devons proposer une compréhension commune de l'interopérabilité qui respecte les critères suivants :

— rendre compte de la complexité de l'interopérabilité, par exemple à travers un modèle en couches et en illustrant les deux composantes de l'interopérabilité (sémantique et technique) ;
— être **exhaustive**, en définissant l'interopérabilité, la démarche d'implémentation, les mécanismes permettant l'interopérabilité et leur caractérisation ;
— être **générique**, pour pouvoir être appliquée à l'ensemble des problèmes d'interopérabilité définis, en considérant les différentes entités et objectifs possibles.

Cette compréhension de l'interopérabilité nous sert à définir le problème de l'interopérabilité des PDRO afin d'assurer un partage d'informations sur les données de recherche et de réduire le coût de recherche de ces données. Pour les deux types de PDRO identifiés, nous devrons définir comment implémenter cette interopérabilité mais aussi extraire les besoins, les verrous et les différents mécanismes associés. Dans le chapitre suivant, nous faisons une proposition de compréhension commune de l'interopérabilité en respectant les critères que nous avons extraits de l'étude de la littérature.

Chapitre 3

Interopérabilité des Plateformes de Données de Recherche Ouvertes

La recherche de données dans la Science Ouverte est décrite comme un processus trop coûteux pour les chercheurs (Van Panhuis et al. (2014); Top et al. (2022); Rainey et al. (2023); Sadeh et al. (2023); European Commission and Directorate-General for Research and Innovation et al. (2021)). Cette recherche de données de recherche nécessite l'exploration de nombreuses Plateformes de Données de Recherche Ouvertes (PDRO) pour pouvoir trouver des données utiles. Pour réduire ce coût, une solution permettant un accès unifié aux métadonnées de ces PDRO est nécessaire. Cet accès unifié nécessite un échange d'informations sur les données de recherche entre les PDRO. Au vu de la grande variété de PDRO et de la haute dissémination des données au sein de ces PDRO (Tanhua et al. (2019)), pour réaliser cet échange d'informations et de données de recherche, ces PDRO doivent être interopérables.

Cependant, la littérature ne propose pas une compréhension de l'interopérabilité des PDRO. L'analyse des travaux sur l'interopérabilité nous a permis d'extraire trois critères à respecter pour une proposition de compréhension de l'interopérabilité :

— rendre compte de la complexité de l'interopérabilité ;
— être **exhaustive** ;
— être **générique**.

Pour répondre au besoin d'interopérabilité des PDRO, nous proposons une compréhension commune de l'interopérabilité (Section 3.1). Dans un premier temps, nous définissons les entités, les données et les informations utiles ainsi que les concepts d'échange de données et d'échange d'informations (Section 3.1.1) applicables à l'ensemble des problèmes d'interopérabilité. Pour l'**exhaustivité** recherchée, cette solution intègre une définition non ambiguë de l'interopérabilité (Section 3.1.2) qui précise les étapes de son implantation ainsi que des outils permettant de spécialiser cette proposition en plusieurs types d'interopérabilité (Section 3.1.3). Elle comporte aussi des mécanismes pour l'implanter ainsi qu'une catégorisation de ces mécanismes (Section 3.1.4) et des outils d'évaluation quantitative de cette implantation (Section 3.1.5). Pour la **généricité** recherchée, nous validons notre proposition de deux façons (Section 3.2) :

— (1) en répondant à l'ensemble des critères d'une théorie formelle de l'interopérabilité encore non définie, sur laquelle tous les travaux sur l'interopérabilité se basent, proposés par Diallo et al. (2011) (Section 3.2.1) ;
— (2) en sélectionnant différents types d'interopérabilité représentatifs des catégories proposées par Maciel et al. (2024) (Section 3.2.2).

Une fois validée, nous appliquons notre proposition à la problématique d'interopérabilité des PDRO. Nous définissons ainsi les verrous à lever pour que le partage de métadonnées entre les PDRO permette de réduire le coût de la recherche d'information (Section 3.3.1). Pour comprendre quel type de solution est nécessaire en fonction de l'état de l'échange de métadonnées dans la Science Ouverte, nous réalisons deux analyses quantitatives (Section 3.3.2) :

— Une analyse quantitative de l'échange de métadonnées dans la Science Ouverte (Section 3.3.2.3) ;
— Une analyse quantitative des performances des outils de mappings automatiques de modèles de métadonnées (Section 3.3.3).

3.1 Proposition de compréhension commune de l'interopérabilité

Dans le chapitre 2, nous avons observé que les différents travaux ne permettent pas d'avoir une compréhension commune de l'interopérabilité. Faute d'une compréhension commune, les travaux relatifs aux problématiques de l'interopérabilité sont incomplets. Pour pallier cette absence de consensus, nous proposons une compréhension de l'interopérabilité **générique**, applicable à l'ensemble des problèmes d'interopérabilité, et **exhaustive**, en intégrant une définition de l'interopérabilité, une démarche d'implantation, des mécanismes pour cette implantation ainsi que des outils d'évaluation de cette implantation. Pour éviter les ambiguïtés et être le plus exhaustif possible, nous utilisons un formalisme mathématique qui s'appuie sur les notations suivantes :

— $g^{\text{L}}_{e_i}$, une grammaire formelle des données et / ou informations gérées par une entité e_i. Cette grammaire définit les règles de définition d'une donnée ou d'une information par e_i associée à la couche L.
— $l^{\text{L}}_{e_i}$, un langage formel généré par $g^{\text{L}}_{e_i}$, défini comme l'ensemble des mots qui respectent les règles de définition de la couche L.
— $\mathcal{P}(S)$, l'ensemble des parties de l'ensemble S, aussi appelé ensemble puissance, qui contient l'ensemble des sous-ensembles de S.
— $\bigcirc$, l'opérateur de composition de fonctions.

Pour illustrer notre proposition de compréhension de l'interopérabilité, nous appliquons ces différents concepts sur un exemple. Nous prendrons un enseignant souhaitant donner un cours magistral à un étudiant.

3.1.1 Entités communicantes, données, informations, échange de données et échange d'information

Nous avons observé dans la littérature une grande variété d'entités à rendre interopérables (cf Chapitre 2). Nous définissons que *l'interopérabilité s'applique sur des entités communicantes, vise la réalisation d'un objectif grâce à un échange de données et un échange d'information.*

Définition 1 - Entité communicante : *Une entité communicante $e \in E_c$ est une entité ayant la capacité d'envoyer des messages et / ou la capacité à recevoir des messages. E_c est l'espace des entités communicantes.*

Il existe deux types d'entités communicantes dans E_c : les **entités communicantes émettrices** et les **entités communicantes réceptrices**. Une entité communicante peut être émettrice et réceptrice. La seule contrainte définissant une entité communicante est sa capacité de communiquer. Ces entités peuvent être d'une grande variété : des personnes, des serveurs, des organisations, des entreprises, des gouvernements, etc...

Dans notre exemple, les entités communicantes sont : un professeur souhaitant transmettre ses connaissances et un étudiant souhaitant recevoir ces connaissances.

Définition 2 - Donnée, information et information utile :

- *Les données* sont les objets nécessaires pour réaliser un objectif visé.
- *Les informations* sont des données décrivant les éléments de la communication (objectif, données à échanger, contexte, etc...).
- *Les informations utiles* sont un sous-ensemble des informations, nécessaires à la communication entre les deux entités communicantes sachant le contexte de la communication.

Reprenons notre exemple. Le professeur souhaite construire un discours permettant de transmettre des connaissances à un étudiant lors d'un cours magistral. Les données sont des phrases contenant des idées. Les informations décrivent les données (les phrases) qui contiennent les idées à échanger mais aussi l'environnement, le contexte, la date, la langue utilisée ou les compétences des personnes de la conversation. Sans définition plus précise de l'objectif ou du contexte, les informations utiles sont uniquement les informations permettant de décrire les données (le sujet, la langue utilisée, la prononciation).

Si nous modifions le médium de communication pour l'enseignement, en passant d'une communication vocale à une communication écrite, nous transformons la prononciation des personnes d'une information utile à une information non nécessaire (à la communication écrite).

Définition 3 - Échange d'informations : *Soit e_1, une entité communicante émettrice, et e_2, une entité communicante réceptrice. L'échange d'informations est le processus de déplacement d'informations de e_1 vers e_2 pour permettre l'utilisation de données.*

Définition 4 - Échange de données : *Soit e_1, une entité communicante émettrice, et e_2, une entité communicante réceptrice. L'échange de données est le processus de déplacement des données de e_1 vers e_2 pour réaliser l'objectif visé.*

Sachant que la communication possède toujours un objectif visé (création de connaissances, discussion, négociation, etc...), cette communication nécessite toujours un échange de données. Il est toujours nécessaire que des informations accompagnent ces données pour permettre leur utilisation. Ainsi, l'échange de données est toujours accompagné d'un échange d'informations.

Dans notre exemple, le professeur est l'entité communicante émettrice (orateur) et l'étudiant est l'entité communicante réceptrice (auditeur). Les échanges sont réalisés du professeur vers l'étudiant dans le cadre d'un cours magistral. L'échange de données est un déplacement de phrases de l'entité émettrice à l'entité réceptrice. L'échange d'informations est un déplacement d'informations décrivant ces phrases (sujet du cours, langue, prononciation, etc...) de l'enseignant vers l'étudiant.

3.1.2 Interopérabilité

L'interopérabilité est un sujet très étudié mais ne possédant pas de définition commune (cf. Chapitre 2). Globalement, les propositions que nous trouvons sur l'interopérabilité possèdent plusieurs limites :

— les décompositions des problèmes de l'interopérabilité en type d'interopérabilité sont différentes selon les domaines d'application des travaux sur l'interopérabilité, rendant impossible l'application en dehors de leur cas d'application ;
— la distinction entre interopérabilité et type d'interopérabilité n'est pas définie ;
— l'implantation de l'interopérabilité et les outils d'implantation ne sont pas définis.

Dans cette section, nous proposons une définition de l'interopérabilité indépendante de tout contexte ou cas d'utilisation.

Nous définissons l'interopérabilité comme *la capacité de deux entités communicantes à travailler en coopération à travers un échange d'information et de données dans le but d'atteindre un objectif.*

Définition 5 - Interopérabilité : *Soit E_c, l'ensemble des entités communicantes, nous définissons Interop, une relation binaire définie sur $E_c \times E_c$ à valeur dans l'espace booléen $\mathbb{B}$. $Interop(e_j, e_i) = 1$ si et seulement si l'entité e_i peut envoyer des données et des informations à e_j et e_j peut recevoir des données et des informations et peut utiliser les informations utiles à la consommation des données pour réaliser l'objectif. On dit que e_j est interopérable avec e_i*

Dans notre exemple, l'étudiant est interopérable avec l'enseignant si l'enseignant peut transmettre une phrase et les informations nécessaires à sa compréhension et si l'étudiant peut recevoir les phrases et les comprendre.

3.1.3 Implantation de l'interopérabilité

Pour comprendre quels sont les verrous qui bloquent la mise en place de l'interopérabilité, nous proposons de définir les étapes de son implantation.

Définition 6 - Les étapes d'implantation de l'interopérabilité : Nous définissons sept étapes dans l'implantation de l'interopérabilité dans un cadre générique. Ces étapes sont inspirées des différents travaux observés dans le chapitre 2 et le modèle OSI [1]. Ces étapes doivent être appliquées à un contexte pour définir les verrous du type d'interopérabilité généré par l'application de ces sept étapes.

La Figure 3.1 illustre ces différentes étapes sous forme d'un modèle en couches, où chaque couche décrit une étape de l'implantation de l'interopérabilité.

Chaque couche nécessite l'implantation de la couche sous-jacente. L'absence de mise en place de la couche L1 implique une absence d'interopérabilité. Chaque couche est une étape dans la réalisation de l'objectif et la mise en place de l'ensemble des étapes assure d'avoir les capacités pour réaliser l'objectif visé.

Nous définissons deux groupes de couches :

— Les couches "système" (couches 1, 2 et 3) répondent à des objectifs techniques et peuvent être assimilées à la réalisation d'une **interopérabilité technique**, décrite dans la littérature. Dans la suite, nous utilisons "interopérabilité technique" pour parler de ce groupe de couches.

1. `https://www.itu.int/rec/T-REC-X.200/fr`

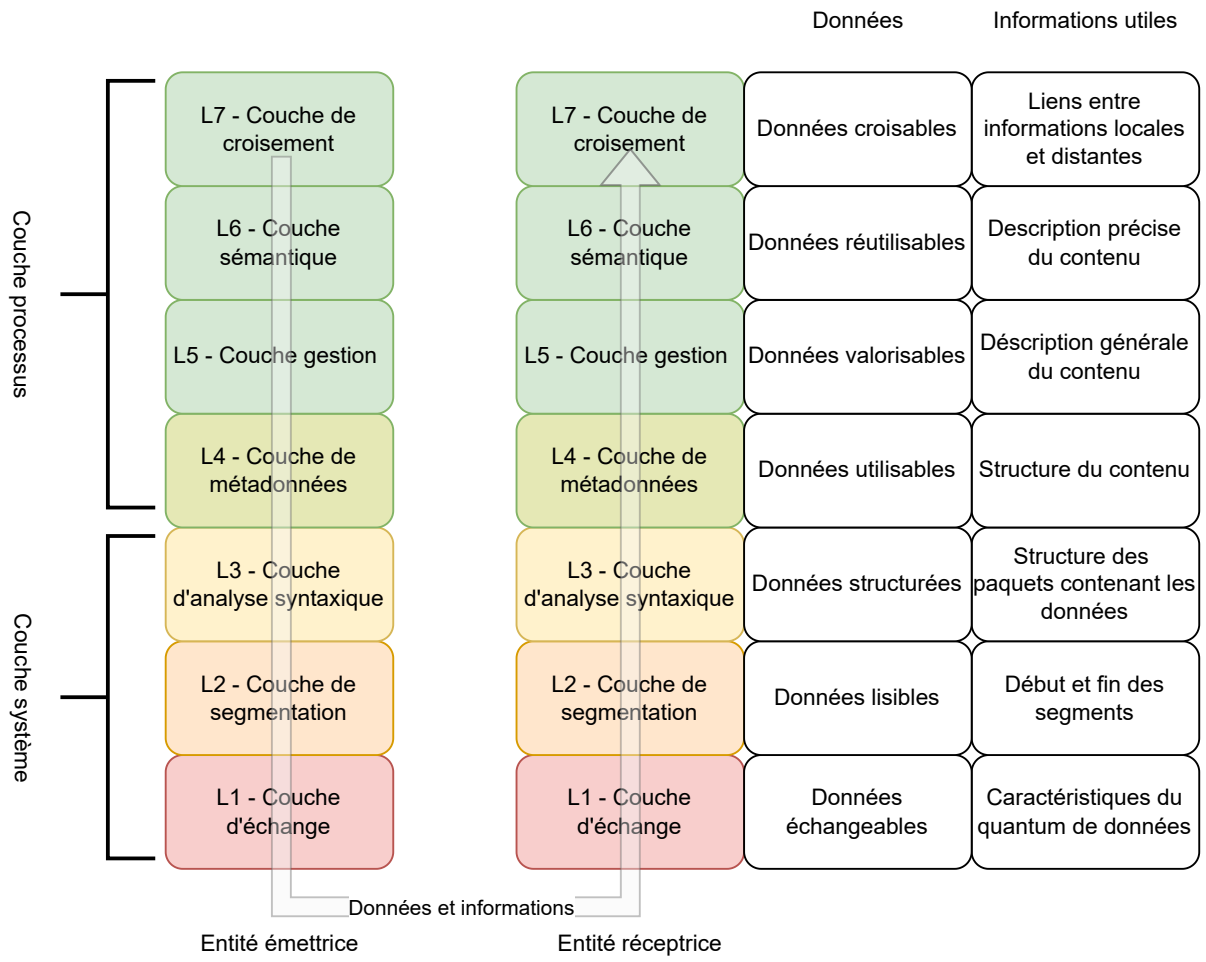

FIGURE 3.1 – Modèle en 7 couches de l'interopérabilité

— Les couches processus (couches 4, 5, 6 et 7) répondent à l'objectif de compréhension des données échangées et peuvent être assimilées à la réalisation d'une **interopérabilité sémantique** décrite dans la littérature. Dans la suite, nous utilisons "interopérabilité sémantique" pour parler de ce groupe de couches.

Pour chaque couche, nous décrivons l'objectif de cette couche et les informations utiles à la réalisation de cet objectif. Les sept couches sont les suivantes :

— Couche L1 - La couche d'échange ("Exchange layer") : Un flux continu de données peut être échangé entre l'entité A (émetteur) et l'entité B (récepteur). La communication entre ces deux entités est possible grâce à la mise en place de mécanismes d'envoi pour l'émetteur et de mécanismes de réception pour le récepteur. Le récepteur reçoit une masse de données constituée de quanta de données, définie comme une donnée brute. À ce stade, les données sont échangeables. **Objectif :** Permettre l'échange d'un flux de données de l'entité émettrice vers l'entité réceptrice. **Informations utiles :** Les caractéristiques du quantum d'information (taille du quantum, les symboles le composant, etc...).

— Couche L2 - La couche de segmentation ("Segmentation layer") : Le flux continu de quantums de données peut être segmenté en tronçons de quantum de données. À ce stade, les données sont lisibles. **Objectif :** Permettre le stockage des données

grâce au découpage du flux de données en tronçons (avec un début et une fin). **Informations utiles :** La connaissance du début et de la fin de chaque segment

— Couche L3 - La couche d'analyse syntaxique ("Parsing layer") : Les tronçons, prenant la forme d'un paquet, peuvent être réorganisés pour reconstituer la structuration initiale du message et le message peut être lu dans son format initial. À ce stade, les données sont structurées. **Objectif :** Permettre de reconstruire la structure des données. **Informations utiles** : La structure et l'organisation des paquets

— Couche L4 - La couche de métadonnées ("Metadata layer") : Le message est décrit par les informations minimales permettant de comprendre le format du message initial et ainsi de lire le contenu du message. À ce stade, les données sont utilisables. **Objectif :** Permettre la lecture du contenu des données basée sur la localisation dans la charge utile des différentes parties de la donnée. **Informations utiles :** Des informations minimales sur la structure du contenu.

— Couche L5 - La couche de gestion ("Management layer") : Les données sont décrites par des informations générales minimales permettant la réutilisation de ces données dans le même contexte que leur utilisation initiale. À ce stade, les données sont valorisables. **Objectif :** Comprendre globalement ce que contient la donnée. **Informations utiles :** Des informations générales (le nom, l'identifiant, le type de données, etc...).

— Couche 6 (L6) - La couche de sémantique ("Semantic layer") : Les données sont décrites en profondeur permettant une compréhension du contenu. À ce stade, les données sont réutilisables. **Objectif :** Comprendre ce que contient précisément la donnée. **Informations utiles :** Des informations précises sur le contenu permettant son utilisation dans un contexte différent du contexte initial (par exemple : grandeur physique mesurée, unité de mesure, etc...).

— Couche 7 (L7) - La couche de croisement ("Crossing layer") : Les informations sur les données peuvent être passées du référentiel de compréhension de l'entité émettrice au référentiel de compréhension de l'entité réceptrice. L'entité réceptrice a ainsi la possibilité de comparer les informations reçues avec des informations locales et permettre le croisement de ces données reçues avec les données locales. À ce stade, les données sont croisables. **Objectif :** Mettre en place des liens entre les données locales et les données échangées. **Informations utiles :** Des liens entre les informations locales et les informations échangées.

Pour chaque couche, un objectif précis est défini et une liste de mécanismes pour la réalisation de l'objectif est associée lors de l'application de ce modèle. Ces mécanismes sont des mécanismes d'interopérabilité et diffèrent selon le contexte de la communication (cf Section 3.1.4).

L'application du modèle d'implantation de l'interopérabilité à un problème spécifique passe par la définition du triplet de l'interopérabilité.

Definition 7 - Le triplet de l'interopérabilité : *Le triplet de l'interopérabilité est un triplet défini sur* $(E_c \times E_c, O, C)$ ***définissant le contexte de la communication****, déterminant les mécanismes d'interopérabilité disponibles pour chaque couche de la mise en place de l'interopérabilité, avec* E_c *l'ensemble des entités communicantes,* O *l'ensemble des objectifs d'une communication,* C *l'ensemble des éléments du contexte influant sur la communication, les mécanismes d'interopérabilité et / ou l'objectif de la couche.*

La spécification des entités communicantes, de l'objectif et d'éventuelles informations

permet la définition des spécificités des solutions. Chaque entité doit être interopérée à l'autre en réalisant l'ensemble des étapes de l'interopérabilité. L'objectif permet de spécifier les objectifs de chaque couche et le contexte définit d'éventuelles contraintes impactant la réalisation de l'objectif (par exemple, le médium de communication, les contraintes environnementales, législatives et/ou techniques, les besoins, etc...)

Dans notre exemple, le triplet de l'interopérabilité se définit comme suit :

— Entités communicantes : un professeur (orateur) et un étudiant (auditeur)
— Objectif : transmettre des connaissances
— Contexte : les deux personnes sont physiquement proches, le volume sonore ambiant est suffisamment bas pour permettre un échange oral

L'objectif est la transmission d'une idée. Cet objectif n'inclut pas la compréhension de l'idée mais uniquement la réception. La compréhension de ces idées nécessiterait que l'étudiant puisse accéder à des connaissances connexes pour réussir à faire le lien avec l'information reçue. Cet enrichissement peut venir du professeur, grâce à un enrichissement de son discours ou par l'étudiant par la récupération de références bibliographiques supplémentaires. Le contexte permet de s'assurer, pour chaque couche, des mécanismes à mettre en place, avec potentiellement la description des mécanismes déjà communs aux deux entités.

Nous illustrons l'application du modèle d'implantation de l'interopérabilité basé sur ce triplet de l'interopérabilité :

— Couche 1 : L'orateur peut faire des sons avec sa voix et l'auditeur peut entendre ces sons avec ses oreilles.
— Couche 2 : L'orateur peut prononcer des syllabes et l'auditeur peut distinguer les différentes syllabes.
— Couche 3 : L'orateur peut prononcer des mots et l'auditeur peut entendre ces mots.
— Couche 4 : L'orateur peut construire des mots en français et l'auditeur peut comprendre les mots en français.
— Couche 5 : L'orateur peut réaliser des phrases simples pour énoncer son idée et l'auditeur peut comprendre l'idée générale de la phrase.
— Couche 6 : L'orateur peut réaliser une succession de phrases complexes qui se suivent pour construire un discours riche et l'auditeur peut extraire de ce discours une idée précise et détaillée.
— Couche 7 : L'auditeur peut extraire les connaissances du discours de l'orateur et les formuler avec ses propres mots.

Pour illustrer l'impact du changement du contexte, supposons que les deux personnes ne sont plus proches physiquement. Le nouveau triplet de l'interopérabilité de notre exemple devient :

— Entités communicantes : un professeur (orateur) et un étudiant B (auditeur)
— Objectif : transmettre des connaisances
— Contexte : les deux personnes sont physiquement éloignées, le volume sonore ambiant est suffisamment bas pour permettre un échange oral

Cet aspect du contexte influe sur les mécanismes de la couche L1, nécessitant la mise en place d'un mécanisme d'interopérabilité pour permettre l'échange de données et d'informations entre les deux personnes.

Définition 8 - Type d'interopérabilité *Un type d'interopérabilité est défini comme l'application du triplet de l'interopérabilité à un problème spécifique.*

Dans notre exemple, le type d'interopérabilité que nous étudions peut être défini

comme "l'interopérabilité des enseignants auprès des étudiants".

Définition 9 - Interopérabilité globale : *Quand l'ensemble des couches du modèle d'implantation de l'interopérabilité sont implantées, l'interopérabilité est dite* ***globale****.*

L'objectif peut être réalisé quand une interopérabilité globale est implantée. Dans notre exemple, l'interopérabilité globale implique que l'élève comprend le discours de l'enseignant et peut construire une réponse à ce discours.

3.1.4 Mécanismes d'interopérabilité

Pour réaliser cette implantation, des mécanismes d'interopérabilité sont nécessaires.

Supposons un triplet de l'interopérabilité fixé. D'un point de vue formel, les données ou informations que peut gérer l'entité e_i respectent les règles de la grammaire g_{e_i} qui contient l'ensemble des règles existantes dans les grammaires de chaque couche (c-à-d l'ensemble des $g^{\mathrm{L}}_{e_i}$). Nous définissons la grammaire de définition des données ou des informations pouvant être gérées par e_i :

$$g_{e_i} = \bigcap_{l=1}^{7} g^{\mathrm{L}}_{e_i}$$

De la même manière, nous définissons l_{e_i}, le langage des données ou des informations pouvant être géré par une entité e_i, généré par la grammaire g_{e_i} :

$$l_{e_i} = \bigcap_{l=1}^{7} l^{\mathrm{L}}_{e_i}$$

Les données ou les informations respectant ces grammaires ou ces langages définissent les données ou les informations qui peuvent être comprises et utilisées par une entité, définissant son **référentiel de compréhension**.

Définition 10 - Mécanisme d'interopérabilité : *Un mécanisme d'interopérabilité est une opération permettant la réalisation de tout ou une partie d'un objectif d'une couche de la mise en place de l'interopérabilité.*

Soit $f^{\mathrm{L}}_{e_i \to e_j}$, un mécanisme d'interopérabilité de la couche L du référentiel de compréhension de l'entité e_i vers le référentiel de compréhension de l'entité e_j. Ces mécanismes peuvent utiliser plusieurs approches pour réaliser l'implantation d'une couche.

Propriété - Type de mécanisme d'interopérabilité : *Les mécanismes d'interopérabilité sont de deux catégories :*

- *Les mécanismes d'interopérabilité par standardisation : Pour une même couche de l'implantation de l'interopérabilité, les deux entités communicantes utilisent le même outil, permettant une interopérabilité native (c-à-d les deux entités possède le même référentiel de compréhension) entre les deux entités. Par exemple : utiliser la même langue (par exemple l'anglais).*
- *Les mécanismes d'interopérabilité par mise en place de passerelle : Un outil d'interopération est mis en place pour permettre le passage du référentiel de compréhension (g_{e_i} ou l_{e_i}) de l'entité émettrice pour celui de l'entité réceptrice. Ces mécanismes peuvent se décomposer en deux types :*

— *Les mécanismes d'interopérabilité s'appliquant sur les grammaires g_{e_i} sont des* ***traducteurs****.*
— *Les mécanismes d'interopérabilité s'appliquant sur les langages l_{e_i} sont des* ***dictionnaires****.*

Dans notre exemple, l'utilisation d'un outil de visioconférence entre le professeur et l'étudiant est un mécanisme d'interopérabilité par standardisation, définissant un médium de communication commun aux deux personnes. Supposons maintenant que le professeur possède un appareil permettant la visioconférence mais pas l'étudiant. Un mécanisme d'interopération par mise en place de passerelle serait de demander à une tierce personne ayant un appareil permettant la visioconférence de transmettre oralement les messages reçus du professeur par visioconférence à l'étudiant.

Définition 11 - Implantation d'une couche : *Une couche est implantée si un ou plusieurs mécanismes d'interopération sont implantés entre les référentiels de compréhension des deux entités permettant la réalisation de l'objectif de ladite couche.*

Supposons e_i une entité communicante émettrice et e_j une entité communicante réceptrice. Une couche est implantée s'il existe $F^{\mathrm{L}}_{e_i \to e_j}$, une composition de mécanismes d'interopérabilité $f^{\mathrm{L}}_{e_i \to e_j}$ pour la couche L, du référentiel de compréhension de l'entité e_i vers le référentiel de l'entité e_j.

$$\text{Une couche est implantée} \Longrightarrow \exists F^{\mathrm{L}}_{e_i \to e_j} : l^{\mathrm{L}}_{e_i} \to l^{\mathrm{L}}_{e_j}, F^{\mathrm{L}}_{e_i \to e_j} = \bigcirc f^{\mathrm{L}}_{e_i \to e_j}$$

Cette définition indique qu'il est nécessaire d'avoir un ou plusieurs mécanismes (par la composition d'un ou plusieurs mécanismes) permettant de passer du référentiel de compréhension de l'entité émettrice au référentiel de compréhension de l'entité réceptrice. Elle n'inclut aucune notion qualitative ou quantitative sur l'implantation d'une couche, ne permettant pas d'illustrer la quantité de données et/ou informations qui pourront être comprises par cette implantation (cf. Section 3.1.5 pour une évaluation de l'implantation).

Dans notre exemple, nous observons que l'utilisation d'outils de visioconférence permet d'implanter la première couche de l'interopérabilité entre un professeur et un étudiant en vue d'une transmission de connaissance à distance.

Pour différencier ces mécanismes, nous proposons de les caractériser selon leur complétude.

Propriété - Complétude d'un mécanisme d'interopération : *Un mécanisme d'interopération $f^{L}_{e_i \to e_j}$ est dit* ***complet*** *s'il permet seul une implantation de l'objectif de la couche en lien.*

Un mécanisme d'interopérabilité $F^{\mathrm{L}}_{e_i \to e_j}$, d'une couche L, du référentiel de compréhension de l'entité e_i vers le référentiel de compréhension de l'entité e_j est dit "complet" si :

$$F^{\mathrm{L}}_{e_i \to e_j} = \bigcirc fi^{\mathrm{L}}_{e_i \to e_j} = f^{\mathrm{L}}_{e_i \to e_j}$$

La complétude des *mécanismes d'interopération par standardisation* est simple à obtenir. Cependant, ce type de mécanisme nécessite de modifier les entités communicantes existantes pour qu'elles intègrent dans le référentiel de compréhension le référentiel de compréhension ciblé par la standardisation. Ce type de mécanisme possède un coût d'adoption élevé, proportionnel au nombre d'entités communicantes à modifier, mais un coût de développement faible. **Les mécanismes d'interopération par standardisation sont adaptés aux propositions de solutions dans un environnement nouveau, sans entité communicante déjà existante.**

La complétude des *mécanismes d'interopération par mise en place de passerelle* est difficile à obtenir. Ce type de mécanisme nécessite de créer des passerelles entre les différents référentiels de compréhension, et la complétude nécessite d'être mise en place pour l'ensemble des référentiels de compréhension à intéropérer. Cependant, les entités communicatrices n'ont pas à être modifiées. Ce type de mécanisme possède un coût de développement élevé, proportionnel au nombre de référentiels de compréhension, mais possède un coût d'adoption faible. **Les mécanismes d'interopération par mise en place de passerelle sont adaptés à la proposition de solutions d'interopération avec des entités communicantes préexistantes.**

Dans notre exemple sur l'interopérabilité d'un professeur et d'un étudiant, l'utilisation d'outil de visioconférence est un mécanisme d'interopération complet car il permet de véhiculer oralement la totalité des données et informations nécessaires. Le besoin de transmettre des documents papiers dans la communication rendrait ce mécanisme d'interopération avec l'utilisation d'outil de visioconférence non complet.

Une autre façon de comparer les mécanismes d'interopérabilité concerne leur capacité à permettre une bi-directionnalité de l'interopérabilité. L'interopérabilité est dite **réciproque** (ou bidirectionnelle) si l'implantation de l'interopérabilité est réalisée dans les deux sens. C'est-à-dire que des mécanismes d'interopérabilité sont implantés entre le référentiel de compréhension d'une entité vers l'autre et inversement.

Définition 12 - Interopérabilité réciproque : *Nous parlons d'interopérabilité réciproque s'il existe* $F^{L}_{e_i \to e_j} : l^{L}_{e_i} \to l^{L}_{e_j}$ *et* $F^{L}_{e_j \to e_i} : l^{L}_{e_j} \to l^{L}_{e_i}$.

Nous définissons $F^{\mathrm{L}}_{e_i \to e_j}$, l'ensemble des mécanismes d'interopérabilité pour la couche L mis en place entre le référentiel de compréhension de l'entité e_i vers le référentiel de compréhension de l'entité e_j et $F^{\mathrm{L}}_{e_j \to e_i}$, une composition de mécanismes d'interopérabilité pour la couche L entre le référentiel de compréhension de l'entité e_j vers le référentiel de compréhension de l'entité e_i.

$$\text{L'interopérabilité est réciproque } \iff \exists F^{\mathrm{L}}_{e_1 \to e_2} : l_{e_1} \to l_{e_2}, F^{\mathrm{L}}_{e_2 \to e_1} : l_{e_2} \to l_{e_1}$$

La réciprocité de l'interopérabilité permet de définir si une communication peut être bidirectionnelle ou non. Dans notre exemple, si l'étudiant peut devenir orateur (par exemple pour poser une question) et que le professeur peut devenir auditeur (par exemple pour entendre et comprendre la question), alors l'interopérabilité est réciproque entre ces deux entités.

Ces mécanismes d'interopérabilité peuvent donc être comparés selon leur complétude et leur bi-directionnalité, en vue de faciliter leur choix. Pour évaluer l'implantation de l'interopérabilité, nous devons définir des outils d'évaluation.

3.1.5 Evaluation de l'implantation

Dans cette section, nous décrivons les outils permettant une évaluation de la qualité d'implantation. Nous proposons d'évaluer quantitativement cette qualité, avec le niveau d'interopérabilité.

Définition 13 - Niveau d'interopérabilité effectif *Le niveau d'interopérabilité effectif entre deux entités est le ratio entre*

— *le nombre de données ou d'informations gérées par une entité communicante émettrice qui sont interopérées par un ou plusieurs mécanismes d'interopération dans le référentiel de compréhension de l'entité communicante réceptrice ;*
— *le nombre total de données ou d'informations* ***gérées*** *par l'entité communicante émettrice.*

Soit $F^{\text{L}}_{e_1 \to e_2} : l^{\text{L}}_{e_1} \to l^{\text{L}}_{e_2}$, une composition de mécanismes d'interopérabilité pour une couche L, du référentiel de compréhension de e_1 vers le référentiel de compréhension de e_2. Soit $Im(F^{\text{L}}_{e_1 \to e_2}) = \{F^{\text{L}}_{e_1 \to e_2}(x) | x \in l^{\text{L}}_{e_1}\}$, le nombre de données ou d'informations gérées par une entité communicante émettrice qui sont interopérées par un ou plusieurs mécanismes d'interopération dans le référentiel de compréhension de l'entité communicante réceptrice. Le niveau d'interopérabilité effectif Nie est défini comme suit :

$$Nie = \frac{|Im(F^{\text{L}}_{e_1 \to e_2})|}{|l^{\text{L}}_{e_1}|}$$

Le niveau d'interopérabilité effectif permet de capturer la quantité d'interopérabilité qui est mise en place entre deux entités pour une couche L. Ce niveau est nécessaire pour **évaluer quantitativement la qualité de l'interopération mise en place entre deux entités**.

Dans notre exemple, le niveau d'interopérabilité effectif est le nombre total de phrases utilisées pour construire une réponse par l'élève sur le nombre de phrases prononcées par l'enseignant. Ce niveau illustre le niveau de compréhension de l'auditeur.

Définition 14 - Niveau d'interopérabilité théorique *Le niveau d'interopérabilité théorique entre deux entités est le ratio entre*

— *le nombre de données ou d'information gérées par une entité communicante émettrice qui sont interopérées par un ou plusieurs mécanismes d'interopération dans le référentiel de compréhension de l'entité communicante réceptrice ;*
— *la quantité totale de données ou d'informations que l'entité émettrice* ***peut gérer****.*

Dans notre exemple, le niveau d'interopérabilité théorique est le nombre total de phrases que pourrait prononcer le professeur et qui pourraient être reformulées correctement par l'étudiant sur le nombre total de phrases que pourrait prononcer le professeur. Ce niveau illustre un niveau de compatibilité intellectuelle entre les deux personnes.

Soit $F^{\text{L}}_{e_1 \to e_2} : l^{\text{L}}_{e_1} \to l^{\text{L}}_{e_2}$, une composition de mécanismes d'interopérabilité entre e_1 et e_2 déployés par e_2 pour implanter une couche L. Soit $Im(F^{\text{L}}_{e_1 \to e_2}) = \{F^{\text{L}}_{e_1 \to e_2}(x) | x \in l^{\text{L}}_{e_1}\}$. Le niveau d'interopérabilité théorique pour une couche Nit est défini comme suit :

$$Nit = \frac{|Im(F^{\text{L}}_{e_1 \to e_2})|}{|l^{\text{L}*}_{e_1}|}$$

avec $*$ l'étoile de Kleene, l'opérateur permettant de définir l'ensemble des mots qui peuvent être inclus dans le langage $l^{\text{L}}_{e_1}$.

Le niveau d'interopérabilité théorique permet de capturer la quantité d'interopérabilité qui pourrait être mise en place entre deux entités, pouvant être interprété dans un langage courant comme le degré de compatibilité de l'entité réceptrice avec l'entité émettrice pour une couche L. Ce niveau est nécessaire pour **évaluer la viabilité d'un projet d'interopération**. Le niveau d'interopérabilité théorique NiT est difficile à quantifier, lié à la difficulté de connaître la fermeture du Kleene des référentiels de compréhension. Cela nécessite de connaître la grammaire formelle définissant le référentiel de compréhension.

Par exemple, la RFC définissant le JSON inclut la définition de la grammaire formelle de construction d'un document JSON [2]. Nous pourrions observer que le niveau d'interopérabilité théorique du format JSON est de 100% avec le format XML, car il existe une transformation de la grammaire d'un format à l'autre sans perte d'information.

Ces deux mesures quantitatives de l'interopérabilité peuvent être calculées pour l'ensemble des couches, en utilisant $F_{e_1 \to e_2}$ à la place de $F^{\text{L}}_{e_1 \to e_2}$, avec $F_{e_1 \to e_2}$ une composition de mécanismes d'interopérabilité permettant de réaliser l'objectif de l'ensemble des couches de l'entité e_1.

3.1.6 Implantation d'un échange d'informations et de données

Une fois l'interopérabilité mise en place, la capacité de ces deux entités communicantes à échanger de l'information et des données est établie. Sans la mise en place de ces mécanismes, l'objectif de l'interopérabilité n'est pas réalisé.

Pour mettre en place un échange d'information, il est nécessaire de mettre en place un mécanisme d'échange d'information et un mécanisme d'échange de données.

Définition 15 - Mécanisme d'échange d'information : *Un mécanisme d'échange d'information est un mécanisme permettant le déplacement d'informations d'une entité communicante à une autre.*

Définition 16 - Mécanisme d'échange de données : *Un mécanisme d'échange de données est un mécanisme permettant le déplacement des données d'une entité communicante à une autre.*

Ces mécanismes d'échange d'informations et de données peuvent être mis en place de trois manières :

— Manuellement ;
— Semi-automatiquement ;
— Automatiquement.

Dans notre exemple, il est nécessaire pour le professeur de prononcer des phrases et pour l'auditeur d'écouter les phrases. Supposons que l'orateur souhaite avoir une connaissance sur un sujet. Nous illustrons chaque type de mécanisme dans notre exemple avec :

— Un mécanisme manuel : l'étudiant se déplace et assiste au cours du professeur.
— Un mécanisme semi-automatique : l'étudiant récupère les notes des autres étudiants du cours.
— Un mécanisme automatique : les cours du professeur sont enregistrés et envoyés directement à l'étudiant.

3.2 Généricité de notre solution

Nous avons proposé une compréhension de l'interopérabilité exhaustive en définissant l'interopérabilité, son implantation, les mécanismes pour cette implantation ainsi que des outils d'évaluation de l'implantation et des mécanismes. Nous souhaitons maintenant valider la généricité de notre proposition. Pour ce faire, nous proposons deux validations :

— Une réponse par notre proposition des critères d'une théorie formelle de l'interopérabilité (cf. Section 3.2.1) ;

2. `https://datatracker.ietf.org/doc/html/rfc8259`

— Une application de notre proposition à une sélection représentative de propositions sur l'interopérabilité afin de les expliquer selon notre proposition (cf. Section 3.2.2).

3.2.1 Evaluation de notre proposition avec les critères d'une théorie formelle de l'interopérabilité

Diallo et al. (2011) ont proposé une analyse de la littérature sur l'interopérabilité pour observer ce à quoi devrait ressembler une théorie formelle de l'interopérabilité sur laquelle se basent tous les travaux sur l'interopérabilité étudiés. Cette analyse a montré que les approches de l'interopérabilité manquent de généricité et d'exhaustivité (Diallo et al. (2011)). Les auteurs décrivent que les termes sont définis de façon ambiguë, amenant à des récursions infinies et que les définitions manquent de formalisation. Ils définissent cinq critères auxquels doit répondre une théorie formelle de l'interopérabilité pour résoudre ces problèmes de compréhension commune.

Ces cinq critères sont :

— **Critère C1** : *Une théorie formelle de l'interopérabilité doit répondre aux critères nécessaires et suffisants de l'interopérabilité. (A formal theory of interoperability should meet the necessary and sufficient requirements for interoperability.)*
— **Critère C2** : *Une théorie de l'interopérabilité devrait être capable de définir formellement une donnée, une information, une information utile et le contexte. (A theory of interoperability should be capable of formally defining data, information, useful information, and context.)*
— **Critère C3** : *Une théorie de l'interopérabilité devrait être capable d'expliquer la dualité de l'interopérabilité. (A theory of interoperability should be capable of explaining the duality of interoperability.)*
— **Critère C4** : *Une théorie de l'interopérabilité doit expliquer l'interopérabilité sans tomber dans une récursion infinie. (A theory of interoperability must be able to explain interoperability without falling into an infinite recursion.)*
— **Critère C5** : *Une théorie de l'interopérabilité devrait être capable d'expliquer l'interopérabilité dans son ensemble. (A theory of interoperability should be able to explain what interoperability is as a whole.)*

Répondre à ces critères permet de s'assurer que notre proposition présente la généricité requise pour définir une théorie formelle de l'interopérabilité. Nous détaillons comment notre cadre de compréhension répond à chacun de ces critères :

Le critère C1 : Le critère C1 nécessite d'intégrer les concepts d'échange d'information et d'utilisabilité des informations. La définition de l'interopérabilité (définition 5) intègre la notion d'échange d'information et d'utilisabilité des données et des informations. De plus, nous avons défini l'échange d'information et de données (définitions 3 et 4) ainsi que les informations utiles et leur rôle selon l'étape de l'implémentation de l'interopérabilité (définition 2). Nous avons distingué interopérabilité et échange d'information et défini que l'interopérabilité était un prérequis pour permettre un échange d'information.

Le critère C2 : La définition 2 précise ce que sont une donnée, une information et une information utile dans le cadre de l'interopérabilité. Ces objets sont définis formellement au travers du référentiel de compréhension et de l'utilisation des grammaires formelles. Pour chaque couche, nous avons défini les informations utiles à la réalisation de l'objectif de chaque couche. Le contexte de la communication est intégré grâce au triplet de

l'interopérabilité (définition 7), prenant en compte l'ensemble des critères impactant les mécanismes d'interopérabilité .

Le critère C3 : La dualité de l'interopérabilité se définit par l'opposition de deux écoles de pensée : l'interopérabilité comme une partie d'un système (l'interopérabilité au sein d'un système) ou l'interopérabilité par rapport aux autres systèmes (l'interopérabilité entre les systèmes). Nos définitions de l'interopérabilité s'appliquent à des entités communicantes. Celles-ci peuvent être incluses au sein d'un système ou être un système. Aucune restriction, autre que les capacités de communication de ces entités, ne limite l'application de notre proposition. Cette approche nous permet de proposer une solution aux deux écoles de pensée.

Le critère C4 : Nous avons défini l'interopérabilité sans récursion. Les termes importants de cette définition ne se basent pas sur l'interopérabilité pour être définis.

Le critère C5 : Nous avons proposé des définitions se basant sur une formalisation mathématique afin de garantir des définitions exhaustives des objets. Notre proposition définit sans ambiguïté le concept d'interopérabilité, mais aussi comment la mettre en place, quels sont les outils à disposition pour cela, et comment l'évaluer. L'interopérabilité est ainsi définie dans son ensemble, avec tous les besoins associés à sa mise en place.

Notre proposition permet de répondre aux différents critères auxquels doit répondre une théorie formelle de l'interopérabilité. Notre proposition peut servir de base pour la conception d'une théorie formelle de l'interopérabilité montrant théoriquement que notre solution est générique. Nous proposons de valider empiriquement la généricité de notre proposition en l'appliquant à une sélection représentative des types d'interopérabilité de la littérature.

3.2.2 Adéquation de notre proposition aux types d'interopérabilité de la littérature

Pour valider empiriquement la généricité de notre solution, nous réalisons une sélection de types d'interopérabilité représentatifs de l'ensemble des catégories de type d'interopérabilité défini par Maciel et al. (2024) (cf. Tableau 3.1). Les auteurs définissent une classification en 5 types d'interopérabilité : les types d'interopérabilité techniques haut et bas niveau, les interopérabilités sociotechniques individuelles et organisationnelles, et les interopérabilités transversales. Par souci de clarté, le tableau 3.1 représente les quatre premières types seulement. La seule interopérabilité transversale de cette revue systématique est l'interopérabilité légale. Nous avons sélectionné des interopérabilités de toutes les types (cf. types d'interopérabilité soulignés dans le tableau 3.1), auxquelles nous ajoutons l'interopérabilité légale. Nous avons sélectionné une proposition de compréhension de l'interopérabilité complète (le modèle LCIM de Tolk et al. (2007)) pour assurer que notre proposition soit applicable à plusieurs types d'interopérabilité.

3.2.2.1 L'interopérabilité légale

L'interopérabilité légale est définie par un rapport technique de la Commission Européenne [3]. L'interopérabilité légale est définie comme l'action de veiller à ce que les organi-

3. https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf

	Technique	Socio-technique		Technique	Socio-technique
Bas niveau / individuel	Technique	Culturelle	Haut niveau/ organisationnel	Pragmatique	Organisationnelles
	Dispositif	Savoir		Sémantique	Entreprises
	Matériel	Règles		Syntaxique	Processus d'entreprises
	Réseaux	Réseaux sociaux		Identité numérique	Operationnelles
	Objets (IoT)			Programmatique	Processus
	Blockchain			Structurelle	Coalition
	Cloud			Conceptuelle	
	Plateformes (IoT)			Constructive	
	Données			Dynamique	
				Ecosystème	
				Informationnelle	
				Service	
				Systèmes logicielles	
				Système	
				Fonctionnelle	

TABLE 3.1 – Classification des types d'interopérabilité par Maciel et al. (2024)

sations opérant dans des cadres juridiques, des politiques et des stratégies différents soient en mesure de travailler ensemble. Les problématiques d'interopérabilité légale se posent au niveau des interactions entre les lois, les normes techniques ou les réglementations de pays différents. Les pays peuvent posséder des systèmes juridiques différents ou des interprétations différentes d'un même problème. L'interopérabilité légale, telle qu'elle est définie, englobe une grande variété de sous-interopérabilités différentes, avec une variation des entités communicantes (par exemple : gouvernements, normes, entreprises, pouvoir législatifs, etc...), une variation des besoins (par exemple : permettre la mise en place de projets intergouvernementaux, gestion de l'imposition de multinationales, etc...) et une variation des contextes (ex : possibilité de changer la loi ou de mettre en place des interdictions, etc...). Ces interopérabilités nécessitent d'avoir des experts définissant clairement le triplet de l'interopérabilité et de définir la liste des hétérogénéités et des outils disponibles pour répondre à l'objectif défini.

Une définition du triplet de l'interopérabilité des normes techniques entre gouvernements pourrait être :

— Entités communicantes : gouvernements
— Objectif : mettre en place des normes techniques compatibles pour permettre des collaborations
— Contexte : possibilité de créer des lois et de définir les normes techniques

La Figure 3.2 définit une vision gros grain des hétérogénéités qui peuvent être rencontrées lors de la mise en place d'une interopérabilité légale. D'un point de vue technique, nous pensons à deux hétérogénéités. La première est la barrière de la langue empêchant les communications entre organismes. La deuxième est sur les différences pouvant exister dans les processus de création de normes techniques, avec notamment les résultats attendus, la forme des documents ou les critères d'évaluation des normes. Au niveau sémantique, les hétérogénéités s'appliquent sur les définitions et les référentiels utilisés pour ces normes, sur les objets ciblés, les caractéristiques voulues et les objectifs. Cette approche permet de rapidement obtenir de potentielles solutions : mettre en place une langue commune, avoir des traducteurs dans les groupes de travail, mettre en place une normalisation des processus de création, utiliser des normes adoptées par les deux entités comme base de travail pour les définitions.

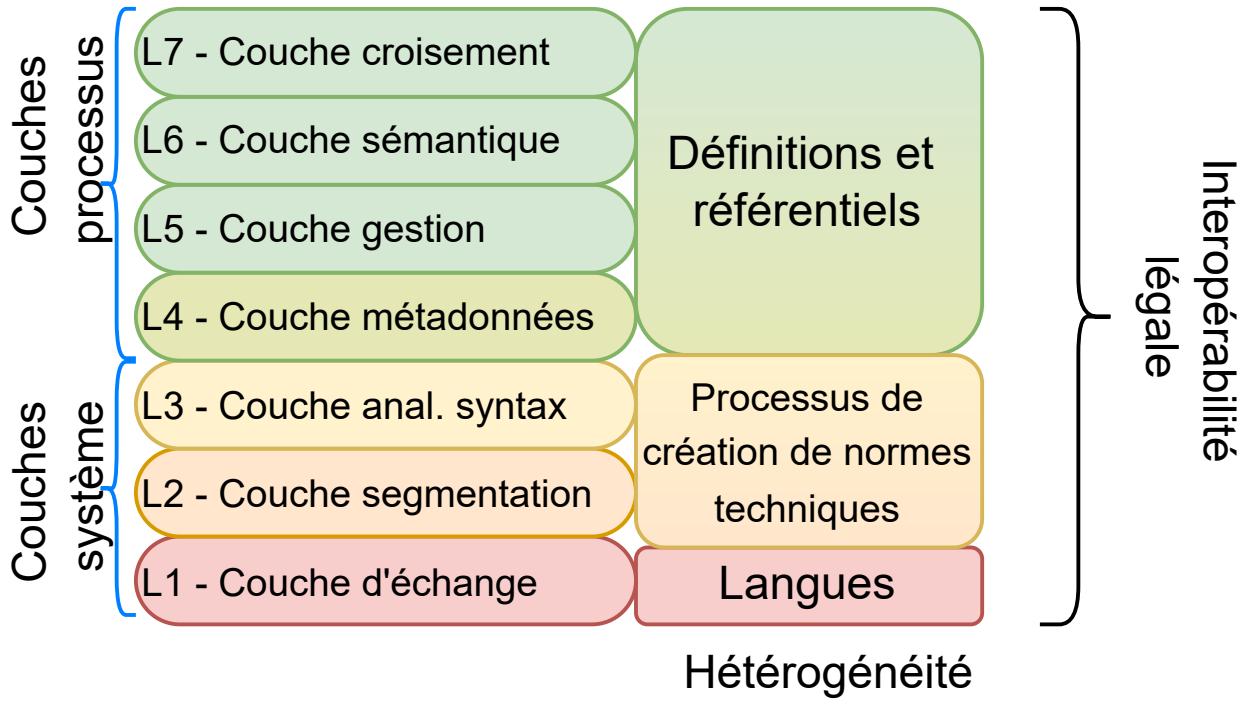

FIGURE 3.2 – L'interopérabilité légale des normes techniques

3.2.2.2 L'interopérabilité du cloud

L'interopérabilité du cloud est définie comme "la capacité des services cloud à travailler en coopération avec d'autres services de cloud et d'autres applications ou plateformes non dépendantes du cloud" (Koussouris et al. (2011)). Ce type d'interopérabilité se place dans le contexte de petites et grandes entreprises utilisant les technologies cloud pour mettre en place des services à valeur ajoutée. Le triplet de l'interopérabilité du cloud se définit comme suit :

— Entités communicantes : plateforme de cloud
— Objectif : coopérations des applications
— Contexte : manque de prise en compte de l'interopérabilité dans le cahier des charges des fournisseurs de cloud

Nous avons défini ce problème avec un triplet de l'interopérabilité unique, se basant sur les plateformes de cloud comme entités communicantes. Le type d'interopérabilité généré par ce triplet est défini dans la figure 3.3.

Une sous-décomposition en plusieurs sous-entités (serveurs, plateforme logicielle de cloud, applications virtualisées) pourrait être proposée pour décomposer ce problème d'interopérabilité et de réduire la complexité de la problématique.

L'interopérabilité du cloud nécessite de répondre aux mêmes problématiques que celles soulevées par le modèle TCP/IP, avec les trois premières couches répondant aux protocoles d'échanges et la dernière couche incluant toutes les problématiques applicatives des solutions de cloud.

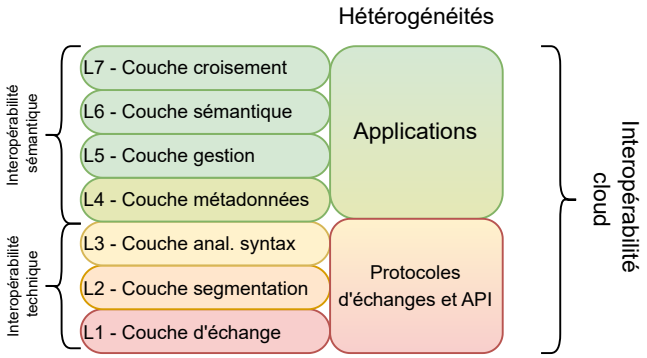

FIGURE 3.3 – Application de notre proposition à l'interopérabilité du cloud (Koussouris et al. (2011))

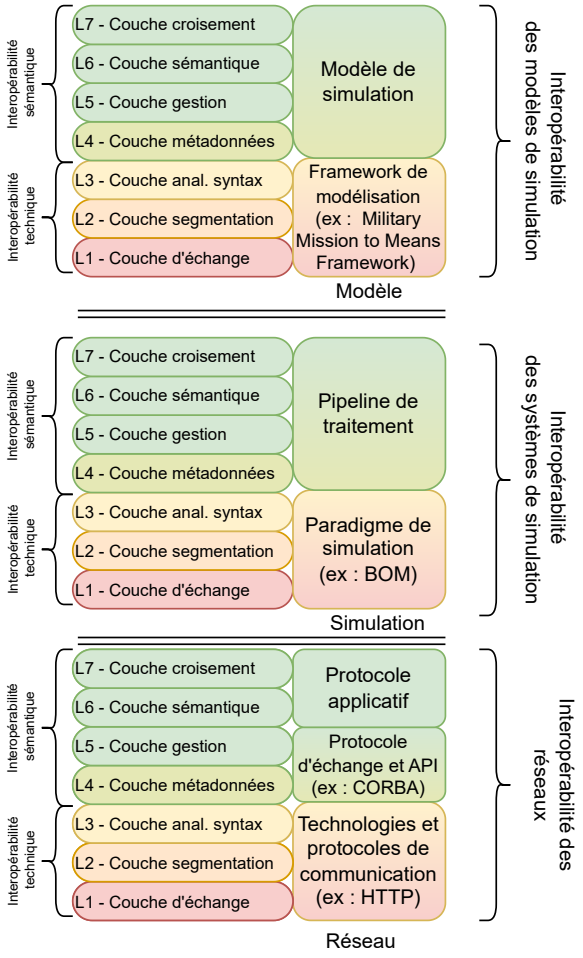

FIGURE 3.4 – Application de notre proposition au modèle LCIM (Wang et al. (2009))

Entité	Notre proposition	LCIM
Modèle	Interopérabilité technique et sémantique des modèles (L1-L7)	Interopérabilité conceptuelle
Système de simulation	Interopérabilité sémantique des systèmes de simulations (L4-L7)	Interopérabilité dynamique
	Interopérabilité technique des systèmes de simulations (L1-L3)	Interoperabilité pragmatique
Réseaux	Interopérabilité sémantique des réseaux (L6-L7)	Interopérabilité sémantique
	Interopérabilité sémantique des réseaux (L4-L5)	Interopérabilité syntaxique
	Interopérabilité technique des réseaux (L1-L3)	Interopérabilité technique

TABLE 3.2 – Comparaison des approches entre le modèle LCIM et notre modèle appliqué

3.2.2.3 Modèle LCIM :

Le modèle “Level of Conceptual Interoperability Model” (Wang et al. (2009)) est un modèle de définition de l’interopérabilité. Ce modèle définit 6 types d’interopérabilités, qui s’organisent dans un modèle en couche pour atteindre la composabilité des systèmes de modélisation et de simulation. Les auteurs définissent que “la modélisation comprend le travail conceptuel de définition du modèle en abstrayant ses données, ses processus et ses contraintes de la réalité, tandis que la simulation se concentre sur la mise en œuvre de ces modèles en tant qu’exécutables, généralement sous la forme de programmes informatiques.”

Les auteurs définissent trois groupes d’entités communicantes à rendre interopérables :
- Les réseaux ;
- Les simulations ;
- Les modèles.

Chaque groupe d’entités possède des objectifs différents, respectivement :
- La mise en place de coopérations entre les solutions applicatives présentes dans les réseaux ;
- La coopération des simulations avec des pipelines de traitements différents ;
- La composition de modèles de simulations.

Ces trois entités, avec leurs objectifs associés, permettent la définition de trois triplets de l’interopérabilité impliquant trois types d’interopérabilité illustrés dans la figure 3.4 : interopérabilité des réseaux, interopérabilité des systèmes de simulation et interopérabilité des modèles de simulation. Nous avons agrégé les couches selon les exemples d’outils utilisés dans l’article, où chaque outil permet d’implanter plusieurs couches de l’interopérabilité. Nous décrivons dans le tableau 3.2 la correspondance entre les couches du modèle LCIM et les étapes d’implantation de notre solution, qui permet d’implanter les couches du modèle LCIM.

Notre proposition permet d’expliquer l’implantation d’un travail entier de compréhension de l’interopérabilité. De plus, nous enrichissons ce travail avec la démarche d’implantation de chaque couche du modèle LCIM.

3.2.2.4 L'interopérabilité culturelle

L'interopérabilité culturelle est définie comme "la mesure dans laquelle les connaissances et les informations sont ancrées dans un modèle unifié de sens à travers les cultures" (Koussouris et al. (2011)). Cette interopérabilité se place dans un monde se transformant en marché unifié et vise à répondre à la problématique de l'échange entre les entreprises sachant les différences culturelles entre régions ou pays.

Le triplet de l'interopérabilité culturelle proposée par Koussouris et al. (2011) se définit comme suit :

— Entités communicantes : entreprises
— Objectifs : mettre en place des collaborations, cogérer et comprendre les processus d'entreprises de régions différentes, échanges commerciaux entre entreprises
— Contexte : régions différentes

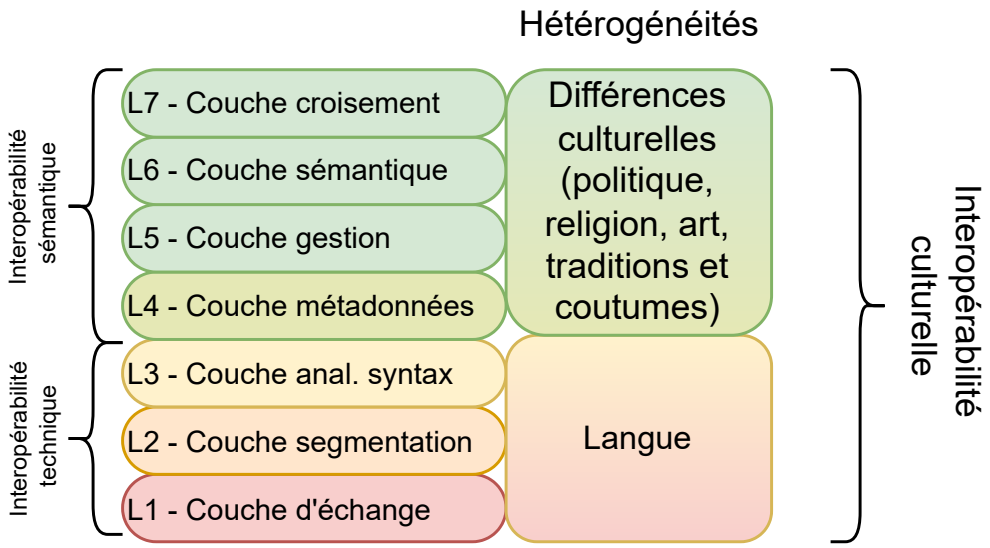

FIGURE 3.5 – L'interopérabilité culturelle (Koussouris et al. (2011))

3.2.2.5 L'interopérabilité organisationnelle

L'interopérabilité organisationnelle est définie comme "la capacité des organisations à communiquer efficacement et échanger des données avec du sens (informations) malgré l'utilisation d'une variété de systèmes d'information sur des types d'infrastructures très différents, éventuellement à travers des régions géographiques et des cultures différentes" (Rezaei et al. (2014b)). Cette interopérabilité est basée sur 3 sortes d'interopérabilités : l'interopérabilité technique, l'interopérabilité syntaxique et l'interopérabilité sémantique. Le triplet de l'interopérabilité organisationnelle se définit comme suit :

— Entités communicantes : organisations
— Objectif : communiquer efficacement et échanger des données avec du sens
— Contexte : variété de systèmes d'information, variété de types d'infrastructures, variété de régions, variété de cultures

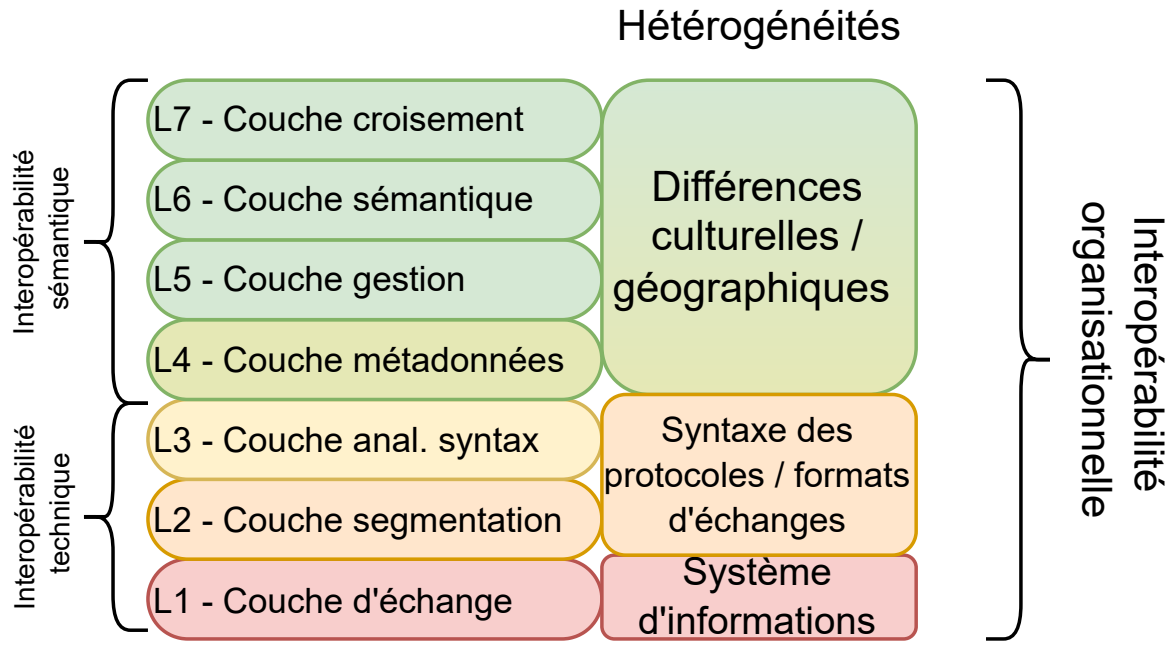

FIGURE 3.6 – L'interopérabilité organisationnelle (Rezaei et al. (2014b))

L'interopérabilité culturelle correspond à la spécification du triplet de l'interopérabilité sur un seul type d'entité : les organisations (cf. Figure 3.6). La décomposition en interopérabilité technique, syntaxique et sémantique correspond à deux problèmes au sein des couches "système" : la composante technique des échanges entre système d'information et la structuration d'une part, et le formatage des messages envoyés entre ces systèmes d'information d'autre part. Les couches "processus" décrivent l'interopérabilité sémantique de l'organisation concernée en tenant compte des différences culturelles et géographiques.

3.2.2.6 Bilan

Nous avons défini notre solution afin qu'elle puisse être appliquée à l'ensemble des catégories de types d'interopérabilité. Ainsi, elle est suffisamment générique pour permettre d'expliquer et de répondre à la grande variété de besoins et d'entités à intéropérer que l'on trouve dans la littérature. Notre proposition permet d'enrichir certaines propositions, notamment en précisant comment implémenter différents types d'interopérabilité définis dans l'état de l'art (cf Section 3.2.2.3).

Notre proposition correspond donc à une compréhension de l'interopérabilité **exhaustive** et **générique**, qui répond à l'ensemble des problématiques de l'interopérabilité. Elle peut servir de base à une théorie formelle et commune, partagée de l'interopérabilité.

3.3 L'interopérabilité des Plateformes de Données de Recherche Ouvertes

Notre proposition de compréhension de l'interopérabilité peut s'appliquer au partage de métadonnées dans la Science Ouverte. Elle peut ainsi permettre l'interopérabilité des PDRO.

En nous appuyant sur cette solution, nous définissons le problème de l'interopérabilité des PDRO dans la section 3.3.1. Pour identifier le type de solution qui y réponde, nous analysons où en est le partage de métadonnées dans la Science Ouverte selon les critères de notre définition. Cette analyse quantitative concerne l'échange d'informations dans la Science Ouverte et nous la présentons dans la section 3.3.2. Elle est basée sur l'étude des mécanismes par standardisation. La mise en place de passerelles sur des API ne pose pas de problème majeur, grâce à la capacité d'abstraction des API qui fournissent un accès transparent à un service quelle que soit la technologie utilisée pour l'implémenter.

Pour compléter cette évaluation quantitative, nous analysons en section 3.3.3 les performances des outils d'alignement (mapping) de modèles de métadonnées en les appliquant à une sélection de ces modèles utilisés dans la Science Ouverte. Ceci nous permet d'évaluer leur adéquation dans ce contexte.

3.3.1 Définition de l'interopérabilité des PDRO

Notre proposition permet de définir le problème d'interopérabilité, ainsi que son implémentation, les différents mécanismes d'interopérabilité et d'évaluation de l'implémentation de l'interopérabilité. Nous pouvons donc répondre à la question initiale. Nous souhaitons rendre interopérables les PDRO pour assurer le partage de métadonnées afin de réaliser une recherche de données unifiée, interdisciplinaire et intercommunautaire. Les entités communicantes sont les PDRO.

Pour définir ce type d'interopérabilité et les étapes de son implémentation, nous devons spécialiser le triplet de l'interopérabilité à notre besoin. Nous définissons ce triplet comme suit :

- Les entités communicantes : les PDRO ;
- L'objectif : mettre en place un partage intracommunautaire, intradisciplinaire, intercommunautaire et interdisciplinaire de métadonnées sur les données de la recherche ;
- Le contexte : Internet, passage par les métadonnées pour le partage des données, grande variété de modèles de métadonnées, grande variété d'API de communication.

Cependant, nous savons que ces PDRO peuvent être décomposées fonctionnellement en deux objets : la mise à disposition d'accès aux données et aux fonctionnalités d'une PDRO et la gestion des métadonnées (cf. Chapitre 2). Résoudre l'interopérabilité de ces deux sous-entités permet de réduire la complexité du problème. Nous proposons donc de décomposer l'interopérabilité des PDRO en deux types d'interopérabilité : l'interopérabilité des API de communication et l'interopérabilité des systèmes de gestion de métadonnées.

Le premier triplet de l'interopérabilité définit l'interopérabilité des API de communication des PDRO :

- Les entités communicantes : les API de communication ;

— L'objectif : mettre en place un partage intracommunautaire, intradisciplinaire, intercommunautaire et interdisciplinaire de métadonnées sur les données de recherche ;
— Le contexte : Internet.

Le second triplet de l'interopérabilité définit l'interopérabilité des systèmes de gestion de métadonnées des PDRO :

— Les entités communicantes : les systèmes de gestion de métadonnées ;
— L'objectif : mettre en place un partage intracommunautaire, intradisciplinaire, intercommunautaire et interdisciplinaire de métadonnées sur les données de recherche ;
— Le contexte : Internet.

L'interopérabilité des API de communication est un prérequis pour permettre l'utilisation des informations et l'interopération des systèmes de métadonnées car l'accès aux systèmes de gestion de métadonnées est réalisé via des API dans les PDRO. Chaque type d'interopérabilité définit un référentiel de compréhension, avec les règles de construction de messages entre les API et les règles de modélisation des informations entre les systèmes de métadonnées. La réalisation de l'objectif global entre les PDRO nécessite d'implémenter une interopérabilité globale de ces PDRO. Cela signifie implémenter une interopérabilité globale des API de communication et une interopérabilité globale des systèmes de gestion de métadonnées.

Nous souhaitons donc implémenter ces deux types d'interopérabilité afin d'assurer un accès unifié aux métadonnées décrivant les données de recherche grâce à un mécanisme d'échange de données et d'informations entre les PDRO. La Figure 3.7 illustre le modèle d'implémentation défini, avec un exemple de deux PDRO que nous souhaitons rendre interopérables, avec des exemples d'outils mis en place pour chaque groupe de couches.

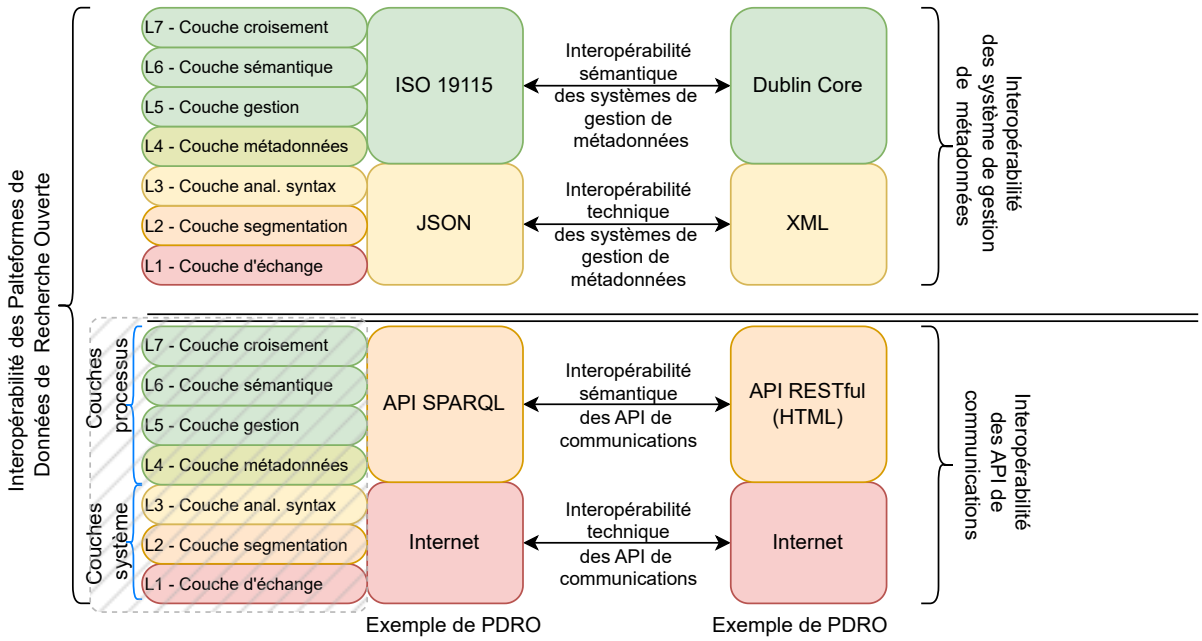

FIGURE 3.7 – L'interopérabilité dans la Science Ouverte

Nous avons noté dans le chapitre 2 que les formats de métadonnées ne posent pas de problème d'interopérabilité majeur, grâce à un grand niveau d'interopérabilité théorique

entre ces formats. Toutes ces PDRO sont déployées sur Internet. L'interopérabilité technique des PDRO est réalisée grâce à la standardisation des protocoles Internet.

Nous souhaitons donc explorer quantitativement l'état de l'échange entre les PDRO de métadonnées sur les données de recherche dans la Science Ouverte. Nous nous basons sur la variété des modèles de métadonnées et la variété des API de communication comme critère d'évaluation.

3.3.2 Évaluation de l'échange de métadonnées dans la Science Ouverte

Nous souhaitons faire un état des lieux de l'échange de métadonnées entre les PDRO afin de trouver quel type de solution doit être conçu (Nosek (2019)). Pour cela, nous mesurons quantitativement l'échange de métadonnées entre les PDRO, en comptant le nombre de PDRO ayant implémenté une interopérabilité globale avec d'autres PDRO ainsi que les PDRO qui échangent des métadonnées. Nous n'avons trouvé aucune analyse de ce type dans la littérature. Pour permettre cette évaluation quantitative et son interprétation, nous calculons la probabilité de trouver une donnée recherchée dans la Science Ouverte.

Nous avons sélectionné le site Re3data.org [4], pour la construction d'un jeu de données afin de réaliser cette évaluation quantitative des métadonnées échangées entre les PDRO. Re3data est un catalogue international de PDRO. Ce site contient des informations sur 3117 PDRO recensées (visité le 31 Mai 2023, cf. Figure 3.8). Ces informations sont utilisées par l'Union Européenne comme indicateur de l'état de la Science Ouverte [5], ce qui permet d'avoir confiance dans leur représentativité. Ce site nous permet d'étudier quantitativement les deux problématiques de l'interopérabilité des PDRO.

3.3.2.1 Variété 1 : Les API de communication

Sur ces 3117 PDRO, 1701 n'ont pas renseigné d'information sur leur API. Pour les PDRO ayant recensé des API, nous observons une moyenne de 1.46 API par PDRO. L'utilisation de plusieurs API permet de répondre à une plus grande variété de besoins, notamment pour que les utilisateurs puissent utiliser les solutions qu'ils connaissent.

Nous observons 3 groupes distincts dans les API utilisées. D'abord, l'API REST se distingue de toutes les autres API comme étant l'API la plus utilisée, par près de 45% des PDRO ayant renseigné des informations sur les API. La proportion d'API REST est similaire à notre analyse comparative dans le chapitre 2 qui confirme que l'API REST est largement répandue dans les PDRO.

Ensuite, un second groupe utilise FTP et OAI-PMH respectivement près de 23% et 20% des PDRO. Enfin, les autres API sont utilisées entre 7% (pour l'API SWORD) et 3% (pour les API SPARQL). Dans les informations, nous observons que des API sont définies en tant que "other". Ces API représentent le second groupe d'API les plus utilisées. Aucune information n'est décrite sur les API qui composent ce groupe. Dans ce contexte, nous considérons que ce groupe peut correspondre à une grande variété de solutions. Nous supposons que ces solutions sont des solutions ad-hoc pour répondre à des problèmes locaux. De ce fait, il n'est pas possible de les utiliser globalement.

4. https://www.re3data.org/

5. 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

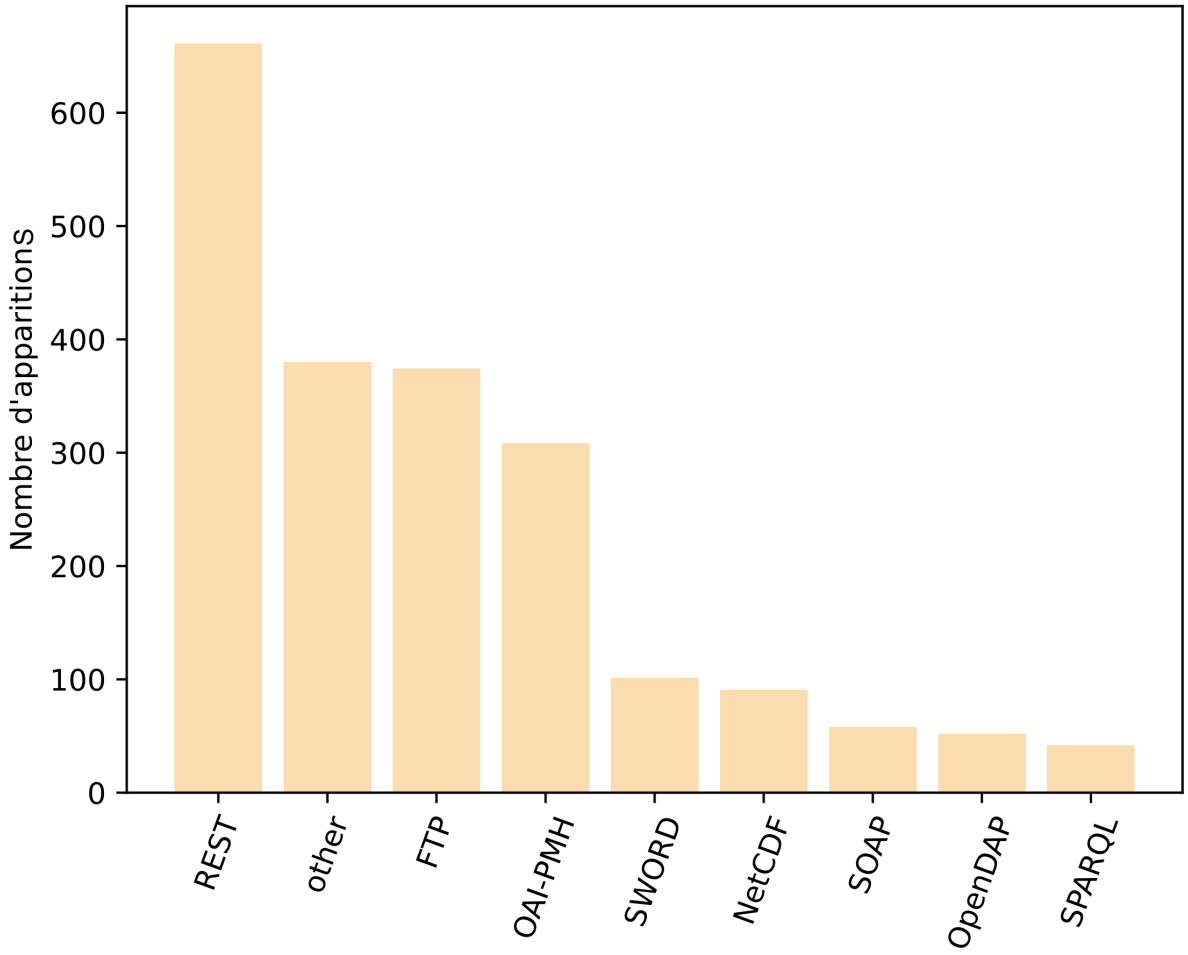

FIGURE 3.8 – Décompte des différentes API recensées dans Re3data.org

L'utilisation d'une API REST pour répondre à cette variété d'API de communication est viable mais ne peut se baser uniquement sur une interopérabilité par standardisation. En effet, le choix d'une API REST permet une interopérabilité par standardisation avec la moitié des PDRO. Ce type d'API permet une intégration d'un grand nombre d'outils et un accès transparent à un grand nombre de technologies. Des passerelles avec les autres types d'API doivent être mises en place pour assurer l'interopérabilité.

3.3.2.2 Variété 2 : Les modèles de métadonnées

Le second niveau d'hétérogénéité observé concerne le contenu des messages avec la modélisation des métadonnées. Les modèles de métadonnées utilisés par ces PDRO sont très variés (cf. Figure 3.9) : nous en avons identifié vingt-neuf différents. Nous comptons une moyenne de 1.48 modèles par PDRO.

Dans ces PDRO, les cinq modèles les plus utilisés (Dublin Core, Datacite, DDI, RDMS, ISO 19115) le sont pour des cas d'utilisation génériques, comme de l'archivage (Datacite) ou de la gestion de données géographiques (ISO 19115). Le modèle le plus utilisé est le modèle Dublin Core. Ce modèle est le seul adapté à OAI-PMH, ce qui peut expliquer sa position de leader sur ces PDRO.

Ces modèles sont des modèles génériques. Les modèles génériques ne peuvent répondre

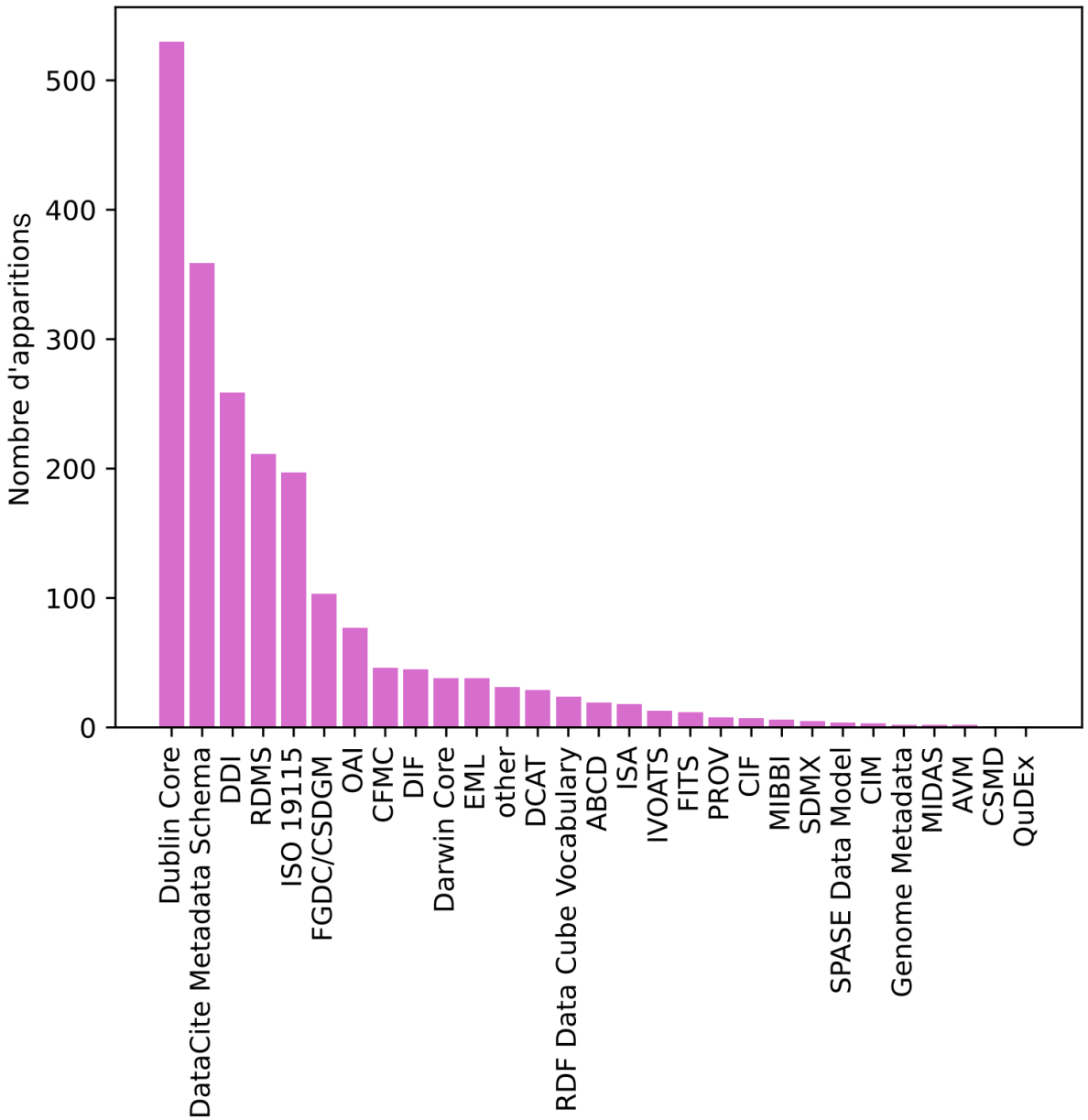

FIGURE 3.9 – Décompte des différents modèles de métadonnées recensés dans Re3data.org

à des besoins spécifiques à des disciplines ou des communautés. L'utilisation d'un modèle spécialisé à des informations disciplinaires est nécessaire. La standardisation n'est pas une solution envisageable pour répondre à cette problématique de PDRO.

3.3.2.3 Analyse quantitative de l'échange de métadonnées

L'observation des données nous a permis de réaliser deux constats :

— L'API REST est partagée par près de la moitié des PDRO, permettant de retenir une approche complexe par standardisation pour la moitié des PDRO et la mise en place de passerelles avec les autres PDRO ;

— Les modèles de métadonnées ne peuvent être interopérés par standardisation à cause de leur variété.

Nous souhaitons donc réaliser une analyse quantitative de ce partage de métadonnées dans la Science Ouverte.

Pour cette évaluation quantitative, nous représentons le réseau d'échange de données la Science Ouverte comme le graphe

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

avec V_{SciO} l'ensemble des sommets du graphe, où chaque sommet est une PDRO et E_{SciO} l'ensemble des arêtes du graphe, où chaque arête représente un échange de métadonnées entre deux PDRO.

Nous avons choisi de définir une PDRO v par les deux composantes suivantes, qui sont celles qui nous intéressent :

$$v = (API, MD)$$

Dans les données de Re3data, le seul protocole d'échange de métadonnées utilisé qui intègre un mécanisme d'échange de métadonnées est le protocole OAI-PMH. Dans la suite, nous définirons qu'une arête existe entre deux nœuds si ces deux PDRO utilisent OAI-PMH et qu'elles possèdent au moins 1 modèle de métadonnées en commun. Le manque d'information sur les échanges pair à pair entre deux PDRO ou les solutions ad-hoc d'échange de métadonnées ne nous permet pas de savoir si d'autres mécanismes d'échange de métadonnées sont présents. Pour pallier ce manque d'information, nous supposons que toutes les PDRO implantant OAI-PMH se connaissent et échangent de l'information.

La visualisation du graphe (cf. Figure 3.10) montre un graphe avec une densité très faible qui illustre le manque d'échange de métadonnées entre ces PDRO.

Chaque arête définit un partage de métadonnées entre deux PDRO. Une très grande majorité des PDRO ne sont pas connectées (2827 PDRO sur les 3117, soit ∼90%, cf. les nœuds gris dans la figure 3.10), c'est-à-dire que ces PDRO ne partagent leurs métadonnées avec aucune PDRO. Ainsi, la recherche des données gérées par ces PDRO doit être réalisée spécifiquement sur chaque PDRO, qui doit comporter une passerelle vers la PDRO ciblée. Ces chiffres confirment le coût élevé de la recherche d'information dans les PDRO.

En revanche, un ensemble de 290 nœuds interconnectés est présent au centre de ce graphe. C'est au sein de cet ensemble de PDRO que nous voulons évaluer l'échange de métadonnées. Pour cela, nous avons extrait automatiquement des *communautés* avec la méthode de Louvain (Blondel et al. (2008)). Ces communautés sont des groupes de PDRO proches, selon les critères évalués (le modèle de métadonnées utilisé et l'API de communication utilisée). Ces communautés ne permettent pas de décrire les communautés organisées autour d'une discipline. Nous avons ainsi mis en évidence trois communautés distinctes représentées en orange, bleu et vert sur le graphe. Une de ces communautés n'est pas du tout connectée aux deux autres. Ces communautés utilisent des modèles de métadonnées distincts et le protocole OAI-PMH. Nous analysons qu'un partage de métadonnées est réalisé dans chaque communauté. La présence de plusieurs communautés distinctes montre que ce partage de données est réalisé entre des communautés mais pas à l'échelle de la Science Ouverte.

De plus, OAI-PMH est un protocole de moissonnage unidirectionnel. La recherche de données doit être réalisée sur le moissonneur. Les moissonnés ne profitent pas de cette possibilité de partage de métadonnées. Cela reflète que le partage de métadonnées dans la communauté OAI-PMH n'est ni homogène ni complètement implémenté.

D'un point de vue global, la densité du graphe est de ∼0.0046, très proche d'un ensemble de nœuds non connectés. Cette visualisation globale tend à indiquer que le partage de métadonnées n'est pas une réalité à l'échelle de la Science Ouverte.

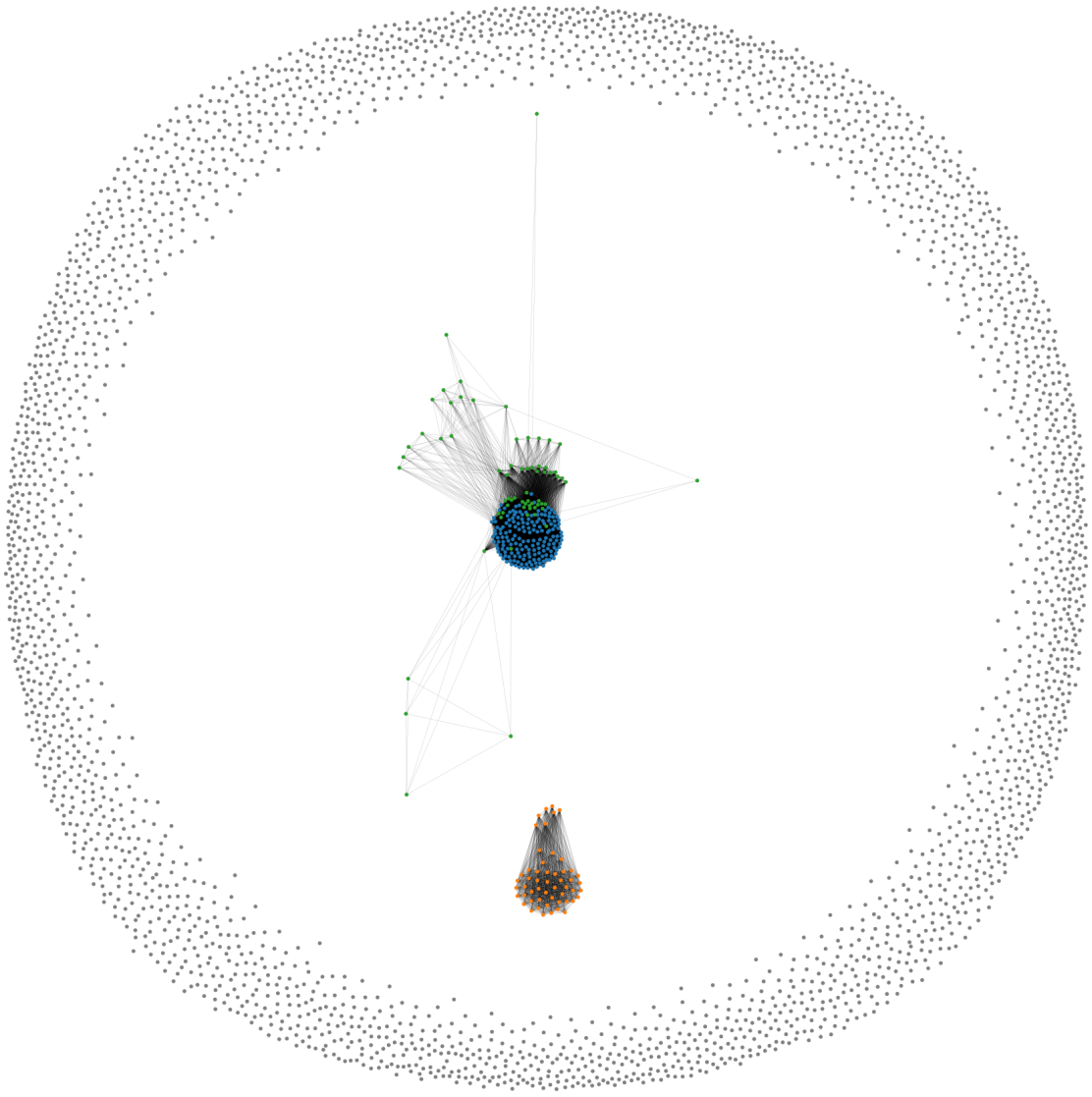

FIGURE 3.10 – Visualisation du graphe G_{SciO} des relations entre PDRO

Pour avoir une évaluation quantitative de cette visualisation, nous définissons un évènement statistique X_d "Trouver la donnée recherchée sur une PDRO". Supposons un chercheur consultant une PDRO choisie aléatoirement dans le graphe. Cet évènement survient lorsque le chercheur recherche un jeu de données spécifique sur une PDRO aléatoire et y trouve les données correspondantes. On définit la probabilité empirique de trouver ces données par la formule suivante :

$$Pr(X_d) = \frac{|dispo(d)|}{|V_{SciO}|}$$

avec $dispo(d)$ l'ensemble des PDRO sur lesquelles les données de recherche d sont stockées (disponibles). Cette probabilité permet d'évaluer le niveau moyen d'échange de métadonnées à l'échelle de la Science Ouverte. OAI-PMH est un protocole qui contient un mécanisme de moissonnage. Les PDRO à moissonner s'inscrivent auprès d'un moissonneur. Il est

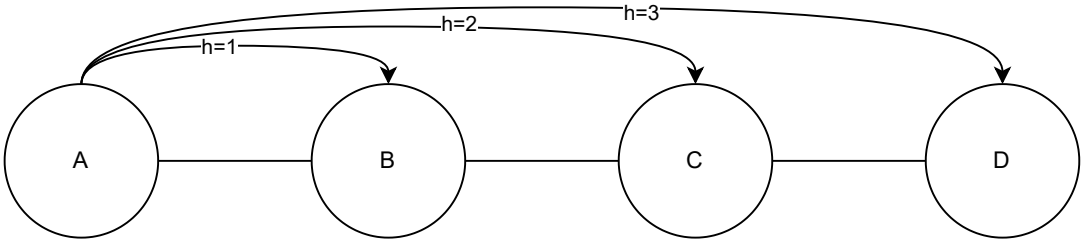

FIGURE 3.11 – Décompte des sauts

possible d'effectuer une recherche d'information sur la PDRO ainsi que pour ses voisins directs virtuellement en même temps. Cependant, il n'est pas possible de rechercher des informations sur les voisins indirects du graphe, faute de mécanisme de propagation du moissonnage. Ainsi, en prenant en compte ce mécanisme, nous utilisons la notion de *nombre de sauts* comme étant le nombre d'arêtes qui séparent deux sommets du graphe. La Figure 3.11 illustre ce comptage de sauts. L'ensemble des PDRO sur lesquelles une donnée d peut être retrouvée devient

$$dispo(d) = \bigcup_{n=0}^{h} S^n(dispo(d))$$

avec

$$S : \mathbb{N} * \mathcal{P}(V_{SciO}) \to \mathcal{P}(V_{SciO})$$

la fonction successeur à saut telle que

$$S^n(sommets) = \underbrace{S \circ ... \circ S(sommets)}_{\text{n fois}}$$

avec $sommets \in \mathcal{P}(V_{SciO})$ et la fonction successeur

$$S : \mathcal{P}(V_{SciO}) \to \mathcal{P}(V_{SciO})$$

retournant l'ensemble des voisins d'un ensemble de nœuds.

Dans le cas d'OAI-PMH, le nombre maximum de sauts est 1. En supposant que les données de recherche ne soient pas dupliquées, la valeur de la fonction $dispo(d)$ est égale au nombre de PDRO qui gèrent ce jeu de données plus le nombre de PDRO voisines que cette PDRO peut moissonner. Au niveau de l'ensemble des PDRO qui proposent OAI-PMH,

$$dispo(d) = 1 + nombre\_moyen\_de\_voisins$$

La distribution du nombre de voisins dans le graphe est visible sur la Figure 3.12. Cette distribution est lissée pour la visualisation. Il existe dans ce graphe un grand nombre de nœuds avec un faible degré, qui diminue de façon exponentielle quand le nombre de degrés augmente. Cette distribution semble suivre une loi de puissance. Le degré maximum est 222, ce qui montre la présence de hubs (nœud possédant un grand nombre de connexions sortantes vers d'autres nœuds) dans l'échange de métadonnées. Cependant, ces hubs ne sont pas reliés à l'ensemble des nœuds ayant un degré supérieur à 1. Cela confirme la présence de communautés distinctes.

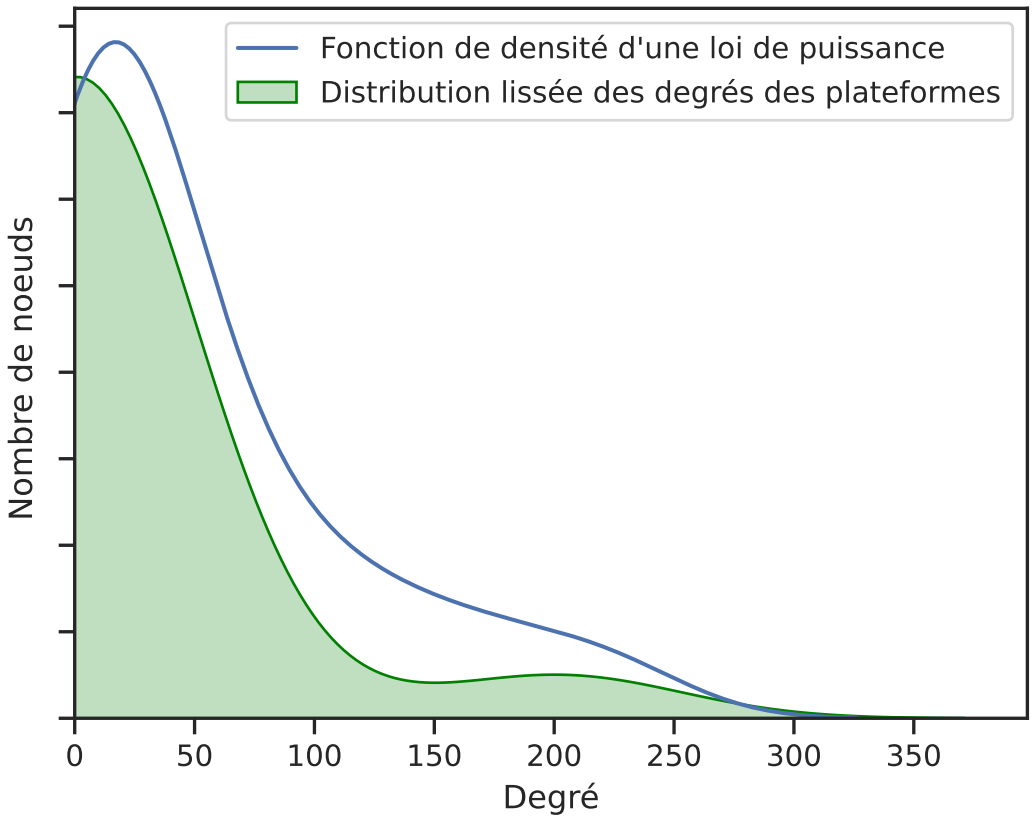

FIGURE 3.12 – Distribution des degrés des nœuds dans G_{SciO}

Il est donc possible d'estimer le niveau moyen d'échange de métadonnées dans la Science Ouverte par la formule suivante :

$$Pr^1(X_d) = \frac{(1 + nombre\_moyen\_de\_voisins)}{|V_{SciO}|}$$

Le nombre moyen de voisins dans le graphe est égal au degré moyen dans le graphe, valant 14.24.

$$Pr^1(X_d) = \frac{1 + 14.24}{3117} \approx 0.5\%$$

La probabilité de trouver une donnée recherchée dans cet échantillon est de 0.5%.

Pour approcher cette probabilité au niveau global de la Science Ouverte, nous avons cherché à estimer un intervalle de confiance du nombre moyen de voisins dans la Science Ouverte. En effet, notre distribution est un échantillon de la distribution réelle des PDRO. Nous supposons que notre échantillon est représentatif de la distribution réelle des PDRO. Nous souhaitons calculer l'intervalle de confiance des voisins moyen. Le niveau de confiance définit la confiance qu'on a sur l'inclusion du nombre moyen de voisins dans cet intervalle.

Le graphe G_{SciO} est un échantillon des PDRO de taille $N = 3117$. L'écart-type des degrés de notre échantillon vaut $S = 5.35$. Pour calculer l'intervalle de confiance du degré moyen de cette distribution, nous avons retenu la formule suivante :

$$nombre\_moyen\_de\_voisins \in [\bar{x} - Z * \frac{S}{\sqrt{N}}; \bar{x} + Z * \frac{S}{\sqrt{N}}]$$

où Z, appelé le Z-score, décrit la position d'une valeur donnée par rapport à la moyenne d'un groupe de valeurs, position mesurée en fonction de l'écart-type, x la variable aléatoire déterminant le nombre moyen de voisins. Cet intervalle de confiance permet de définir, avec un niveau de confiance donné lié à la valeur de Z, que la valeur réelle du paramètre que nous estimons est contenue dans cet intervalle.

Nous prenons une approche utilisée dans les processus de contrôle qualité et appelée "6 sigma". Cette approche impose d'avoir un taux de confiance de 99.9997%, que nous pouvons interpréter comme "il y a 99.9997% de chance que la méthode que nous appliquons produise un intervalle contenant la valeur à approximer". Nous avons donc

$$nombre\_moyen\_de\_voisins \in [14.24 - 5.35; 14.24 + 5.35] = [8.89; 19.58]$$

Donc la probabilité de trouver une donnée recherchée dans l'ensemble des PDRO (i.e. représentatif du niveau d'échange de métadonnées dans la Science Ouverte) se trouve, avec une confiance à 99.9997% dans l'intervalle

$$Pr(X_d) \in [\frac{1 + 8.89}{3117}; \frac{1 + 19.59}{3117}] = [0.003; 0.007]$$

Ainsi, l'échange de métadonnées dans la Science Ouverte permet de n'avoir une probabilité empirique maximale que de 0.7% de trouver la donnée voulue. Nous avons une confiance à 99.9997% que le niveau d'échange de métadonnées soit en dessous de 0.7%. Ce niveau d'échange de métadonnées est **très faible**. Nosek (2019) définit les étapes à mettre en place pour l'implémentation de la Science Ouverte. Le premier niveau d'implémentation est la proposition d'une solution architecturale visant la mise en place de l'objectif. Ce niveau très faible indique que le partage de métadonnées n'est pas présent à l'échelle de la Science Ouverte. Nous devons donc proposer une implémentation du partage de métadonnées avec une solution à un niveau architectural, c'est-à-dire rendre possible l'échange de métadonnées.

3.3.2.4 Limites

Le catalogue de données Re3data ne permet pas d'obtenir les informations sur les échanges de données effectifs. En effet, il se peut que des échanges de métadonnées aient lieu grâce à des solutions ad-hoc occasionnant des collaborations entre PDRO, échanges non renseignés dans ce catalogue. Pour compenser ce manque d'information, nous avons surévalué le niveau d'échanges de métadonnées sur deux points.

Le protocole OAI-PMH possède une grande limite qui est son fonctionnement unidirectionnel. Les échanges d'informations vont des PDRO moissonnées vers les moissonneurs. Les PDRO moissonnées n'ont pas connaissance des informations stockées sur le moissonneur. Dans notre évaluation, nous avons considéré les échanges d'informations bidirectionnels.

De plus, selon le protocole OAI-PMH, pour que le catalogue d'une PDRO soit récolté par un moissonneur, il est nécessaire que les deux PDRO aient connaissance l'une de l'autre, et pour cela que chaque PDRO s'inscrive auprès de l'autre qui est alors considérée comme un moissonneur. Parmi la grande variété de PDRO existantes, il est très peu plausible que toutes les PDRO utilisant OAI-PMH se connaissent. Or, nous avons supposé que toutes les PDRO disposant d'OAI-PMH se connaissaient.

De cette manière, nous avons essayé de pallier le manque d'information potentiel de notre jeu de données. En s'ajoutant au niveau de confiance que nous avons fixé pour l'étude statistique, nous avons une confiance élevée dans notre analyse quantitative.

Notre analyse s'est portée sur l'évaluation de l'échange de métadonnées en observant les mécanismes d'interopérabilité par standardisation. Pour confirmer notre analyse, nous souhaitons évaluer l'autre approche d'interopération par la mise en place de passerelles. Nous avons noté la pertinence des API REST pour la mise en place de passerelles entre des solutions technologiques. Nous souhaitons donc maintenant évaluer les différents outils d'interopérabilité qui facilitent la mise en place de passerelles entre modèles de métadonnées.

3.3.3 Analyse des outils de mise en place de passerelles

Nous venons de rappeler que la mise en place de passerelles entre les API de communication est viable, notamment grâce aux API REST fournissant des accès transparents à de nombreuses technologies (cf. Chapitre 2). Parmi les autres outils d'interopérabilité par mise en place de passerelles sur les modèles de métadonnées, nous étudions la viabilité des mappings (alignements) entre modèles.

Ces mappings peuvent être réalisés de façon automatique, par logiciels de mapping qui implémentent des algorithmes d'alignement d'ontologies ou de données, ou de façon manuelle, grâce à des correspondances (crosswalks) écrites par une ou plusieurs personnes, en général des expertes de ces modèles. Cependant, l'écriture des crosswalks est fastidieuse et coûteuse en temps, ce qui empêche de retenir cette approche comme solution unique, du fait du problème de passage à l'échelle. Nous avons donc souhaité évaluer de manière expérimentale les performances des outils de mappings de l'état de l'art sur plusieurs modèles de métadonnées de la Science Ouverte utilisés sur différentes plate-formes.

3.3.3.1 Sélection de modèles de métadonnées

Pour notre analyse, nous avons sélectionné plusieurs modèles de métadonnées. Tout d'abord, nous avons choisi deux PDRO : ODATIS [6] - une plate-forme de gestion de données océanographiques - et AERIS [7] - une plate-forme de gestion de données atmosphériques. Ces PDRO sont des pôles de données du programme Data-Terra [8] qui vise à gérer les données du système Terre. Ces PDRO proviennent de la même discipline (Sciences naturelles) et de la même communauté (la communauté Data-terra). Ces PDRO ne permettent pas d'accéder simplement au modèle de métadonnées qu'elles utilisent. Pour obtenir ces modèles de métadonnées, nous avons téléchargé des exemples de fichiers de métadonnées provenant de ces PDRO puis nous en avons extrait les modèles.

Ensuite, nous avons sélectionné deux autres modèles de métadonnées définis par un organisme de standardisation, HL7. Le modèle C-CDA vise la modélisation des données médicales et permet un échange de données entre les systèmes d'information. Le modèle FHIR est le successeur de C-CDA avec l'objectif de rendre l'utilisation de ce modèle plus simple que C-CDA. Nous avons récupéré des exemples d'instances de métadonnées de C-CDA [9] depuis un Github contenant des exemples d'implémentations de C-CDA et de

6. https://www.odatis-ocean.fr/
7. https://www.aeris-data.fr/
8. https://www.data-terra.org/
9. github.com/jmandel/sample_ccdas

FHIR[10] depuis sa documentation officielle. À partir de ces exemples, nous avons extrait les modèles de la même manière. Ces quatre modèles nous serviront à évaluer les capacités d'interopération des outils de mappings automatiques.

Chacun de ces modèles a été considéré comme un arbre et stocké sous forme d'une liste de chemins permettant de définir et d'accéder à un concept, avec une racine commune. Chaque chemin est une concaténation des différents niveaux du chemin avec une notation en point (“.”). La figure 3.13 illustre un exemple d'un chemin vers un concept dans le modèle de la plateforme ODATIS. Dans cet exemple, nous n'avons pris qu'une sous-partie du chemin par souci de clarté. De cette manière, la structure du modèle est conservée.

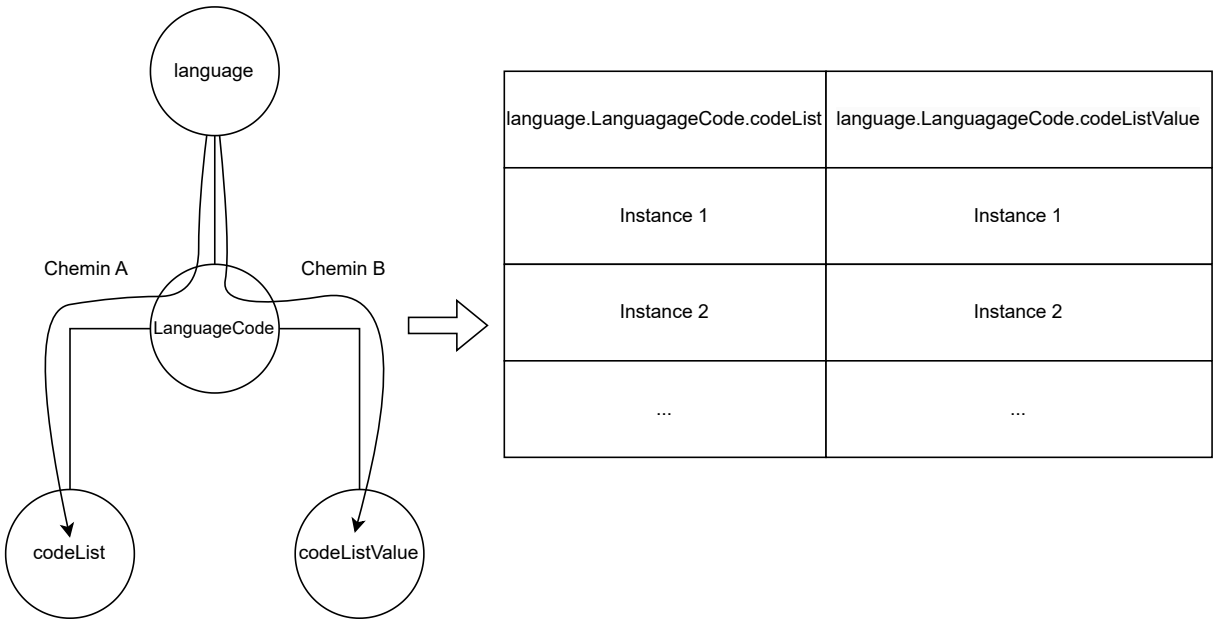

FIGURE 3.13 – Exemple de chemin d'un modèle de métadonnées représenté sous forme de graphe et avec la notation en point.

3.3.3.2 Outils et protocole

Pour l'évaluation des outils de mappings, nous avons focalisé notre sélection sur des logiciels qui ont pour objectif d'aligner des modèles sans considération de leur discipline d'application, qui soient donc génériques et adaptés à l'ensemble des modèles de métadonnées de la Science Ouverte.

Nous avons sélectionné un outil d'alignement de modèles conceptuels évalué dans le cadre de la campagne OAEI[11], qui est la campagne de référence de test des logiciels d'alignement d'ontologies et de graphes de connaissances : COMA ++ (Massmann et al. (2006)). Cet outil obtient un score de F-measure compris entre 0.6 et 1 sur les jeux de données utilisés pour la comparaison des logiciels de mapping au sein d'OAEI, montrant de bonnes à d'excellentes performances selon les cas. De plus, COMA++ est accessible en ligne[12] ce qui permet sa réutilisation et son évaluation. Il est annoncé comme l'outil adoptant la meilleure approche pour la découverte et l'intégration de données dans la

10. www.hl7.org/fhir/index.html
11. https://oaei.ontologymatching.org/
12. https://github.com/delftdata/valentine

librairie Valentine (Koutras et al. (2021)) que nous utilisons pour cette expérimentation. Enfin, COMA++ prend en compte des aspects structurels des modèles sans qu'aucun prétraitement soit nécessaire. Nous souhaitions aussi évaluer un outil de mappings prenant en

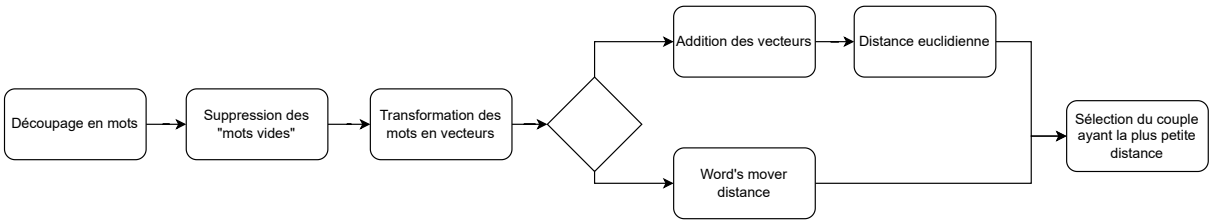

FIGURE 3.14 – Etapes de fonctionnement de notre outil de mappings basé sur les word-embedding

compte des aspects sémantiques des modèles. Des travaux ont été proposés sur le mapping à l'aide de plongements vectoriels de mots (word-embeddings) (Oh et al. (2024); Gonçalves et al. (2019)) mais nous n'avons pas trouvé de proposition qui ne soit pas spécifique à un domaine particulier, qui ouvre le code de l'outil proposé ou qui soit facilement réutilisable. Nous avons donc développé un outil de mappings basé sur des algorithmes de représentation vectorielle de mots ou word-embedding. Le processus de mapping de cet outil est le suivant (cf. Figure 3.14) :

- (1) Découpage des mots du chemin de l'ensemble des concepts des deux modèles ;
- (2) Suppression des mots vides (les mots communs n'apportant pas de sens au corpus, comme par exemple les déterminants) ;
- (3) Transformation de l'ensemble des mots en vecteurs ;
- (4) Application d'une distance sur les vecteurs des deux modèles à aligner ;
- (5) Pour chaque concept d'un modèle, nous conservons la plus petite distance avec un autre concept de l'autre modèle.

Nous obtenons en sortie, pour chaque concept du modèle, un concept associé qui est défini comme égal ou similaire à un concept de l'autre modèle.

Nous avons sélectionné deux distances :

- La distance euclidienne, pour une approche naïve en prenant la somme des mots composant un chemin vers un concept.
- La Word's mover distance, pour une approche considérant le chemin comme étant un corpus afin d'essayer de saisir les liens entre les mots du chemin.

Pour ces outils, une étape de prétraitement des modèles est nécessaire (cf. Figure 3.14). Un découpage des chemins des modèles en mots est réalisé en plusieurs étapes :

- Un découpage des mots, en suivant le format "camel-case" (c'est-à-dire un découpage des mots en fonction des majuscules dans les chaînes de caractères) ;
- Un découpage statistique, basé sur la distribution des mots dans la langue anglaise et découpant selon le mot le plus probable d'apparaître dans une chaîne de caractères.

Une fois ce découpage réalisé, nous supprimons les "mots vides", les mots qui n'apportent pas de sens, tels que les déterminants.

Chaque mot est transformé en vecteur grâce aux modèles de word-embedding (GloVe et Word2Vec).

- En utilisant la distance euclidienne, la somme des vecteurs doit être effectuée et la distance est calculée sur la somme totale des vecteurs des mots du chemin ;

— En utilisant la word's mover distance, la distance est appliquée directement sur la liste des vecteurs du chemin, considérant le chemin comme un corpus afin d'obtenir les potentielles interactions sémantiques entre les mots du chemin.

Pour chaque concept d'un modèle, le chemin de l'autre modèle avec la plus petite distance (pouvant être défini comme le chemin le plus proche selon la distance utilisée) est conservé.

En plus de l'adéquation des outils automatiques de mapping aux modèles de métadonnées, nous souhaitons observer les impacts des choix de modélisation sur l'interopération des modèles. Notre première hypothèse est que plus les modèles définissent les concepts avec une profondeur élevée dans l'arbre du modèle (c'est-à-dire qu'ils utilisent une grande combinaison de concepts différents pour définir un concept particulier) plus la mise en place d'une interopérabilité entre ces modèles est compliquée. Pour illustrer, nous supposons que la définition de l'adresse d'une personne utilisant "adressePersonne" pour la modélisation de cette information est plus facile à rendre interopérable qu'un modèle ayant choisi "personne.information.personnelle.adresse". Notre seconde hypothèse est sur le choix du vocabulaire utilisé. Nous souhaitons évaluer l'impact de la complexité du vocabulaire utilisé pour définir les concepts.

Pour valider ou invalider notre première hypothèse, nous avons sélectionné des sous-chemins dans les chemins de modèles afin de simuler des choix de modélisation différents. Pour chaque chemin, cette sélection est effectuée en partant de la feuille et nous avons augmenté la taille des chemins considérés jusqu'à conserver le chemin complet. Grâce à la variation de la taille de ces chemins, nous avons pu augmenter notre jeu de données et atteindre 631 combinaisons de couples de modèles. Grâce à cette sélection, nous définissons une liste de métriques après avoir sélectionné les sous-chemins (taille minimale du chemin dans le modèle, taille maximale du chemin, taille du modèle, etc...).

Pour valider ou invalider notre seconde hypothèse, nous définissons des métriques de complexité du vocabulaire. Ces métriques se basent sur la distribution des mots de la langue anglaise basée sur le corpus Wikipédia en anglais [13]. Nous supposons que plus la fréquence d'un mot est élevée dans ce corpus, plus il est général et simple, et inversement, moins un mot est fréquent et plus il est spécialisé et complexe. Nous mettons en place une étude des corrélations entre ces métriques et les métriques de performance des outils.

Pour évaluer ces deux outils de mappings (COMA++ et notre développement utilisant des word-embeddings), nous avons associé les modèles de métadonnées en paires pour obtenir des scénarios de mapping différents. Nous avons réalisé un mapping manuel entre ces paires de modèles, qui nous sert de vérité terrain. Ces paires, et le nombre de mappings réalisés, sont :

— ODATIS - AERIS afin d'observer la capacité à calculer des mappings intradisciplinaires (Science naturelle / Géosciences) et provenant de plateformes d'une même communauté (36 mappings dans la vérité terrain) ;
— C-CDA - FHIR afin d'observer la capacité à calculer des mappings entre modèles annoncés comme interopérables (286 mappings dans la vérité terrain) ;
— AERIS - FHIR afin d'observer la capacité à calculer des mappings interdisciplinaires et intercommunautaires (13 mappings manuels).

L'ensemble du code est disponible sous forme de notebooks réutilisables et réexécutables. Le code, les données, les résultats et les vérités terrains sont accessibles sur un dépôt Git [14]. Cette analyse a fait l'objet d'une publication dans la conférence internationale RCIS 2023

13. `github.com/IlyaSemenov/wikipedia-word-frequency`
14. `https://github.com/vincentnam/OS_data_interop_RCIS_2023`

(Dang et al. (2023a)).

3.3.3.3 Résultats

Le tableau 3.3 synthétise les résultats de nos évaluations. Chaque cellule du tableau contient la valeur minimale et la valeur maximale de la métrique correspondante parmi toutes les applications des outils sur les tailles de chemin sélectionnées pour la paire de modèle correspondante.

Les métriques utilisées sont la précision - pourcentage de mappings corrects parmi le nombre total de mappings réalisés, le rappel - pourcentage de mappings qui sont effectivement réalisés parmi ceux qui devraient être réalisés - et le F1-score - qui agrège la précision et le rappel et indique la performance générale de l'outil évalué.

		Paire de modèles	Précision	Rappel	F1-score
COMA++		ODATIS/AERIS	0.00 - **0.27**	0.00 - 0.2	0.00 - **0.21**
		FHIR/C-CDA	0.00 - 0.20	0.00 - 0.04	0.00 - 0.06
		AERIS/FHIR	0.05 - 0.10	0.25 - 0.42	0.08 - 0.16
word-embedding	Distance euclidienne	ODATIS/AERIS	0.00 - 0.09	0.00 - 0.09	0.00 - 0.07
		FHIR/C-CDA	>0 - 0.05	>0 - 0.05	>0 - 0.04
		AERIS/FHIR	0.01 - 0.05	0.08 - **0.50**	0.01 - 0.10
	Word mover distance	ODATIS/AERIS	0.00 - 0.15	0.00 - 0.26	0.00 - 0.19
		FHIR/C-CDA	>0 - 0.09	0.01 - 0.09	0.01 - 0.07
		AERIS/FHIR	0.02 - 0.05	0.25 - **0.50**	0.04 - 0.10

TABLE 3.3 – Résultats de l'alignement des 3 paires de modèles

Nous observons que l'outil qui atteint les meilleurs scores est COMA++, avec un F1-score de 0.21. Cependant, ce score n'est pas suffisant pour assurer une fiabilité des mappings réalisés entre modèles et montre une inadéquation des outils évalués à l'interopération des modèles de métadonnées de la Science Ouverte.

Un résultat surprenant de cette évaluation concerne les mappings du couple C-CDA / FHIR. Ces modèles sont annoncés comme interopérables mais ces résultats semblent contredire cette affirmation, illustrant la difficulté de comprendre ce concept d'interopérabilité. Nous avions souligné ce problème à l'issue de l'étude de l'état de l'art (cf. Chapitre 2).

De plus, nous observons aussi une différence notable entre notre évaluation et celle proposée dans le cadre d'OAEI. Nous expliquons cette différence par le manque de représentation des caractéristiques des modèles de la Science Ouverte dans les campagnes d'évaluation d'OAEI. Les modèles utilisés pour OAEI sont des modèles standardisés et nettoyés. Les modèles de la Science Ouverte possèdent des qualités d'implémentation variables. De plus, les PDRO utilisent des modèles de métadonnées spécifiquement conçus pour répondre aux besoins applicatifs et plus rarement des modèles standardisés. Leurs choix conceptuels sont différents de ceux des organismes spécialisés dans la modélisation de métadonnées.

Pour essayer de caractériser l'inadéquation des outils d'alignement aux modèles de métadonnées de la Science Ouverte et valider notre hypothèse sur la taille des modèles (qui impacte négativement leur interopérabilité), nous avons mis en place plusieurs métriques et observé leurs corrélations aux métriques habituelles de performance des modèles. Ces métriques se basent sur deux hypothèses :

— La taille des modèles impacte négativement les métriques de performance des modèles.
— La spécificité et la complexité du vocabulaire impactent négativement les métriques de performance des modèles.

Corrélation de Spearman	Précision	Rappel	F1-score
	Métriques de taille des modèles		
Quantité de concepts stockés	-0.34 / -0.28	- 0.41 / -0.46	**-0.45 / -0.44**
Nombre de nœuds intermédiaires	-0.27 / -0.34	-0.31 / -0.52	**-0.32 / -0.49**
Ordre (Nombre de nœuds)	-0.31 / -0.33	-0.36 / -0.51	**-0.38 / -0.49**
Taille du vocabulaire	-0.35 / -0.20	-0.44 / -0.03	**-0.46 / -0.15**
Taille minimum de chemin	0.21 / -0.31	0.20 / -0.58	0.28 / -0.52
Taille maximum de chemin	-0.13 / -0.26	-0.06 / -0.45	**-0.08 / -0.42**
	Métriques de complexité du vocabulaire		
Moyenne des fréquences de termes	0.16 / -0.14	-0.19 / -0.18	0.00 / -0.22
Moyenne pondérée des fréquences de termes	0.13 / -0.18	-0.24 / -0.22	-0.03 / -0.25

TABLE 3.4 – Corrélation de Spearman entre les métriques de performance et les métriques sur les modèles

Nous avons défini six métriques sur la taille des modèles :
- La quantité de concepts définis, c'est-à-dire le nombre de lignes ;
- Le nombre de nœuds intermédiaires, c'est-à-dire une approche de la profondeur des concepts ;
- L'ordre, c'est-à-dire le nombre total de sommets dans l'arbre du modèle ;
- La taille du vocabulaire, c'est-à-dire le nombre de mots différents ;
- La taille maximum et la taille minimum des chemins.

Ensuite, nous avons défini deux métriques de complexité du vocabulaire. Nous calculons la fréquence de chaque mot dans le corpus Wikipedia en anglais [15]. Nous nous basons sur l'hypothèse qu'un mot est d'autant plus complexe ou spécialisé que sa fréquence d'apparition est faible. Nous définissons comme métriques : (1) la moyenne des fréquences des termes et (2) la moyenne des fréquences de termes pondérée par leur fréquence dans le vocabulaire.

Le tableau 3.4 reprend nos résultats de calcul de corrélation de Spearman entre les métriques liées aux modèles et les métriques de performance des outils sur les modèles associés. Chaque cellule contient deux valeurs. La première valeur est la corrélation pour le premier modèle de chaque paire (ODATIS, FHIR, AERIS). La seconde valeur est la corrélation pour le second modèle de chaque paire (AERIS, C-CDA, FHIR). Toutes les corrélations possèdent des P-values < 0.05, à l'exception de la moyenne pondérée des fréquences de termes.

QUANTITÉ DE CONCEPTS STOCKÉS / NOMBRE DE NŒUDS INTERMÉDIAIRES /

15. github.com/IlyaSemenov/wikipedia-word-frequency

Ordre du graphe(nombre de nœuds) / Taille du vocabulaire : Les valeurs de corrélation (respectivement (-0.45 / -0.44), (-0.32 / -0.49), (-0.38 / -0.49) et (-0.46 / -0.15)) montre **un impact négatif significatif**, avec des valeurs de p-value ¿ 0.05.

La taille minimum des chemins : Les valeurs des corrélations sont opposées entre celles du premier modèle de la paire et du second modèle de la paire. **Aucune conclusion** ne peut être tirée de ces résultats.

La taille maximum des chemins : Les corrélations sont négatives. Les résultats montrent une *tendance négative* de la taille maximum des chemins sur les performances des outils mais il est difficile de connaître l'impact véritable de cette tendance au vu de la différence entre les valeurs de corrélation selon les tailles des chemins sélectionnés.

Métriques de complexité du vocabulaire : La moyenne des fréquences de termes possède une corrélation nulle pour les premiers modèles de chaque paire. Cependant, les seconds modèles possèdent une valeur de corrélation négative. La spécificité pourrait être liée aux problèmes de performance des outils mais *aucune conclusion* ne peut être tirée de ces résultats.

3.3.3.4 Discussion

Nous avons constaté les limites à l'échange de métadonnées au sein de la Science Ouverte, en explorant la mise en place d'interopérabilité par standardisation. Pour pallier le manque d'information sur une potentielle utilisation des mécanismes d'interopérabilité par mise en place de passerelle, nous avons réalisé une expérimentation permettant d'observer les performances de deux logiciels d'alignement de modèles de métadonnées : COMA ++, un outil validé dans la campagne d'évaluation d'outil d'alignement OAEI, avec un score compris entre 0.6 et 1 de f1-score dans ces campagnes, montrant des performances élevées ; un outil que nous avons conçu, basé sur l'utilisation de word-embeddings, qui sont utilisés dans des outils de mappings dans la littérature, mais que nous n'avons pas pu réutiliser. Nous avons sélectionné quatre modèles de métadonnées utilisés dans la Science Ouverte sur lesquels nous avons évalué nos outils. Nous avons en plus évalué deux hypothèses sur les choix conceptuels de modèles pouvant impacter les performances des outils : (i) la croissance de la taille des modèles a un impact négatif sur les performances des outils de mappings et (ii) l'augmentation de la complexité du vocabulaire utilisé pour définir les modèles impacte négativement les performances des outils de mappings.

Nos résultats ont montré que les outils de mappings évalués ne sont pas adaptés à la mise en place d'interopérabilité entre les modèles de métadonnées utilisés dans la Science Ouverte. Le meilleur outil est COMA++ avec un F1-score de 0.21. Ce résultat est très inférieur aux performances observées dans les campagnes d'évaluation d'OAEI, montrant une différence significative des modèles utilisés dans la Science Ouverte et des modèles utilisés dans ces campagnes d'évaluation.

La taille des modèles impacte négativement les performances des outils. Nous n'avons pas pu obtenir de résultats significatifs sur la complexité du vocabulaire, et n'avons donc pas pu valider ou invalider cette hypothèse. L'interopération des modèles de métadonnées par la mise en place de passerelle automatique (c'est-à-dire à l'aide d'outils d'alignement automatique) n'est pas envisageable avec les performances observées.

Nous avons atteint notre objectif et observé le manque d'adéquation des outils de mappings aux modèles de métadonnées de la Science Ouverte. Cependant, les résultats nous permettent d'envisager trois pistes d'amélioration de l'adéquation de ces outils.

Modèles de métadonnées : La réduction de la taille des modèles n'est pas une solution viable. Les modèles de métadonnées déjà utilisés ne peuvent être modifiés sans un coût élevé. De plus, les choix de modélisation effectués sont *a priori* faits pour une modélisation fine et précise des concepts. La modification des modèles ne semble pas une approche possible dans le contexte de la Science Ouverte. Cependant, nos résultats ont montré que les mappings correctement réalisés sont des mappings sur des métadonnées génériques ou techniques (nom, identifiants, etc...). Ainsi, une catégorisation en 3 groupes de métadonnées (générales, techniques et domaine) comme étape de prétraitement au calcul des mappings permettrait d'avoir une taille réduite des modèles et (sans doute) une augmentation des performances des outils.

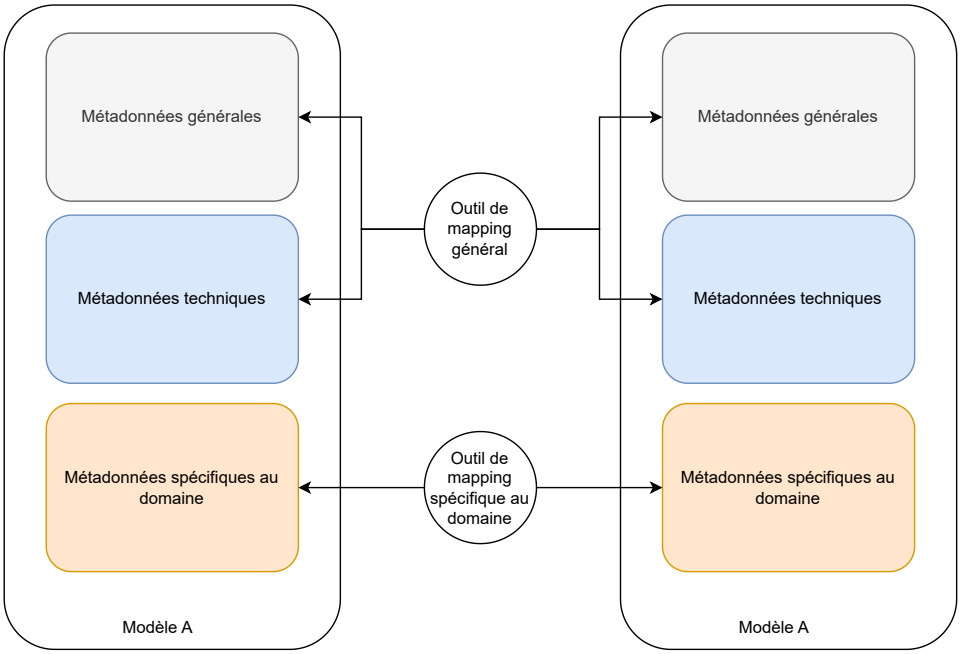

FIGURE 3.15 – Pistes pour améliorer les performances des outils de mappings

Les outils de mappings : Des outils spécialisés à des disciplines et des domaines particuliers ou des types de métadonnées spécifiques montrent des performances meilleures pour ces domaines, notamment la solution de mapping de données biomédicales proposée par Gonçalves et al. (2019) ou la solution de raisonnement basé sur des mappings spécialisés sur les graphes temporels par Li et al. (2021). Les outils que nous avons utilisés sont génériques et mal adaptés à notre objectif. Cependant, ils ont réussi à correctement associer des concepts généraux ou techniques. Combiner l'utilisation d'outils généraux et d'outils spécifiques aux domaines pourrait permettre de meilleures performances de mappings de ces modèles (cf. Figure 3.15). Cependant, cet enrichissement implique une augmentation du coût pour la mise en place de mappings.

Les architectures de gestions de données : La dernière piste d'amélioration concerne les mécanismes manuels d'interopération par mise en place de passerelle. L'écriture manuelle de mappings ne passe pas à l'échelle. Cependant, ces mappings, réalisés par des personnes ayant des connaissances expertes sur les modèles, ont une qualité qui permet d'en avoir une confiance plus élevée. Une solution architecturale peut permettre de proposer une solution de passage à l'échelle de ces mappings manuels. Cette solution doit

être décentralisée, fédérée et distribuée pour permettre à toutes les entités de recherche de participer à la mise en place d'une interopérabilité des modèles. Cet effort communautaire assure un équilibre des charges à l'ensemble des entités de recherche participant au processus et donc un passage à l'échelle des mappings manuels.

Nous avons observé que l'état de l'implémentation de l'échange de métadonnées dans la Science Ouverte nécessite une solution architecturale, c'est-à-dire rendre possible l'échange de métadonnées et une recherche unifiée de données. Nous explorerons dans la suite la troisième piste d'amélioration, basée sur de nouvelles infrastructures, afin de répondre au manque d'adéquation des outils de mappings automatique aux modèles de métadonnées de la Science Ouverte.

3.4 Conclusions

La recherche de données dans la Science Ouverte est décrite comme trop coûteuse par les chercheurs. Pour réduire ce coût, une interopérabilité des Plateformes de Données de Recherche Ouvertes (PDRO) est nécessaire afin de réaliser un partage de métadonnées sur les données de recherche. Nous avons observé dans le chapitre 2 un manque de compréhension commune sur l'interopérabilité ainsi qu'une absence de proposition de compréhension de l'interopérabilité des PDRO.

Nous avons réalisé une proposition de compréhension de l'interopérabilité **générique** et **exhaustive**. L'exhaustivité de notre solution est réalisée par la réponse à la définition de l'interopérabilité, des étapes d'implémetation, des mécanismes pour implanter cette interopérabilité ainsi que des outils d'évaluation de cette implémentation. Nous avons validé la généricité de cette solution en :

— Répondant à l'ensemble des critères nécessaires d'une théorie formelle de l'interopérabilité proposés par Diallo et al. (2011). Cette théorie est définie comme celle sur laquelle tous les travaux de l'interopérabilité se basent.

— Appliquant notre proposition à une sélection représentative de types et de travaux d'interopérabilité, basée sur une catégorisation réalisée par Maciel et al. (2024) et permettant d'expliquer ces différents problèmes au sein d'une unique proposition de compréhension.

L'application de notre proposition a permis d'enrichir les travaux étudiés avec :

- un apport sur les étapes d'implémentation des différents types d'interopérabilité définis dans ces travaux ;
- la définition et la caractérisation des mécanismes d'interopérabilité.

Une fois validée, nous avons appliqué notre proposition à l'interopérabilité des PDRO. Cette application a permis d'extraire deux grands composants pour la mise en place d'une recherche de données unifiée : les outils d'accès aux services, aux données de recherche et aux métadonnées avec les API de communication ainsi que les systèmes de gestion de métadonnées. Nous avons pu identifier deux types d'interopérabilité des PDRO : l'interopérabilité des API de communication et l'interopérabilité des systèmes de gestion de métadonnées. Nous avons observé dans le chapitre 2 que ces deux types d'interopérabilité se heurtent à deux variétés : (1) la **variété des API de communication** et (2) la **variété des modèles de métadonnées**. Nous avons donc proposé une analyse quantitive de l'échange de métadonnées entre les PDRO. Nous avons observé que le niveau d'échange de métadonnées est inférieur à 0.7%, avec une confiance à 99.9997%. Cette analyse se base sur un mécanisme d'interopérabilité par standardisation. Pour s'assurer que ce partage de métadonnées ne soit pas réalisé avec une autre approche, nous avons évalué

les outils de mappings automatiques de modèles de métadonnées, nécessaires pour gérer la variété des modèles de métadonnées. Or ces outils ne sont pas suffisamment performants pour permettre la mise en place d'une interopérabilité des PDRO pour la modélisation des métadonnées.

Faute d'un partage de métadonnées entre les PDRO, nous concluons que l'état de l'échange de métadonnées entre les PDRO indique un besoin de rendre possible cet échange (cf. Nosek (2019)). Nous nous attacherons à proposer une solution permettant de le mettre en place pour fournir un accès unifié aux métadonnées. Cette solution doit s'attacher à proposer une réponse à la variété des API de communication et à la variété des modèles de métadonnées.

Ces travaux ont fait l'objet de deux publications scientifiques dans la conférence RCIS 2023 (Dang et al. (2023a)) et la conférence RCIS 2024 (Dang et al. (2024a)).

Chapitre 4

Recherche intracommunautaire : Lac de Données de la Science Ouverte (LDSO)

La recherche intracommunautaire de données permet aux entités de recherche (chercheur, équipe de recherche ou laboratoire de recherche) de trouver des données de recherche (jeux de données ou articles scientifiques) sur des sujets en lien avec leur discipline. Pour les communautés intrinsèquement interdisciplinaires comme l'astrobiologie (Aydinoglu et al. (2014)), la recherche de données intracommunautaire permet toutefois de rechercher des données interdisciplinaires. Ainsi, les entités de recherche - par la réutilisation des données de recherche trouvées, potentiellement traitées et nettoyées - peuvent répondre à de nouvelles questions de recherche avec par exemple (1) l'entraînement de modèles d'intelligence artificielle qui nécessitent un grand volume de données, (2) la jointure de jeux de données de recherche pour réaliser des analyses croisées ou (3) la recherche d'articles scientifiques pour la mise en place d'une revue systématique de questions.

La multiplication des PDRO et des modèles de métadonnées utilisés complexifie la recherche de données de recherche car elle empêche leur exploration exhaustive. Sans partage intracommunautaire de métadonnées, ces données de recherche, essentielles au travail d'une entité de recherche qui en a besoin, peuvent ne pas être trouvées. Pour illustrer ce constat, nous reprenons notre entité de recherche ER ("Equipe de recherche"). ER souhaite répondre à une nouvelle question de recherche et trouver des données externes à son entité pour la traiter. Pour cela, ER connaît cinq Plateformes de Données de Recherche Ouverte (PDRO) différentes qui sont utilisées dans sa communauté. Pour chacune, ER doit répéter ce même processus (cinq fois) :

— Accéder à la PDRO
— Apprendre à utiliser les outils/interfaces de cette PDRO
— Comprendre la modélisation des métadonnées des données de la plateforme associée
— Rechercher et trier les résultats de la recherche
— Télécharger les données.

Nous illustrons cet exemple par la Figure 4.1. L'entité de recherche ER doit envoyer des requêtes à chaque PDRO, avec des API de communication différentes (cf. les flèches pleines dans la Figure 4.1) et obtient des métadonnées modélisées avec des modèles différents (cf. les flèches en pointillé dans la Figure 4.1). Ces différents API et modèles de métadonnées doivent être appris et maîtrisés par ER. Cet apprentissage d'un coût élevé peut dissuader de réaliser cette recherche de données.

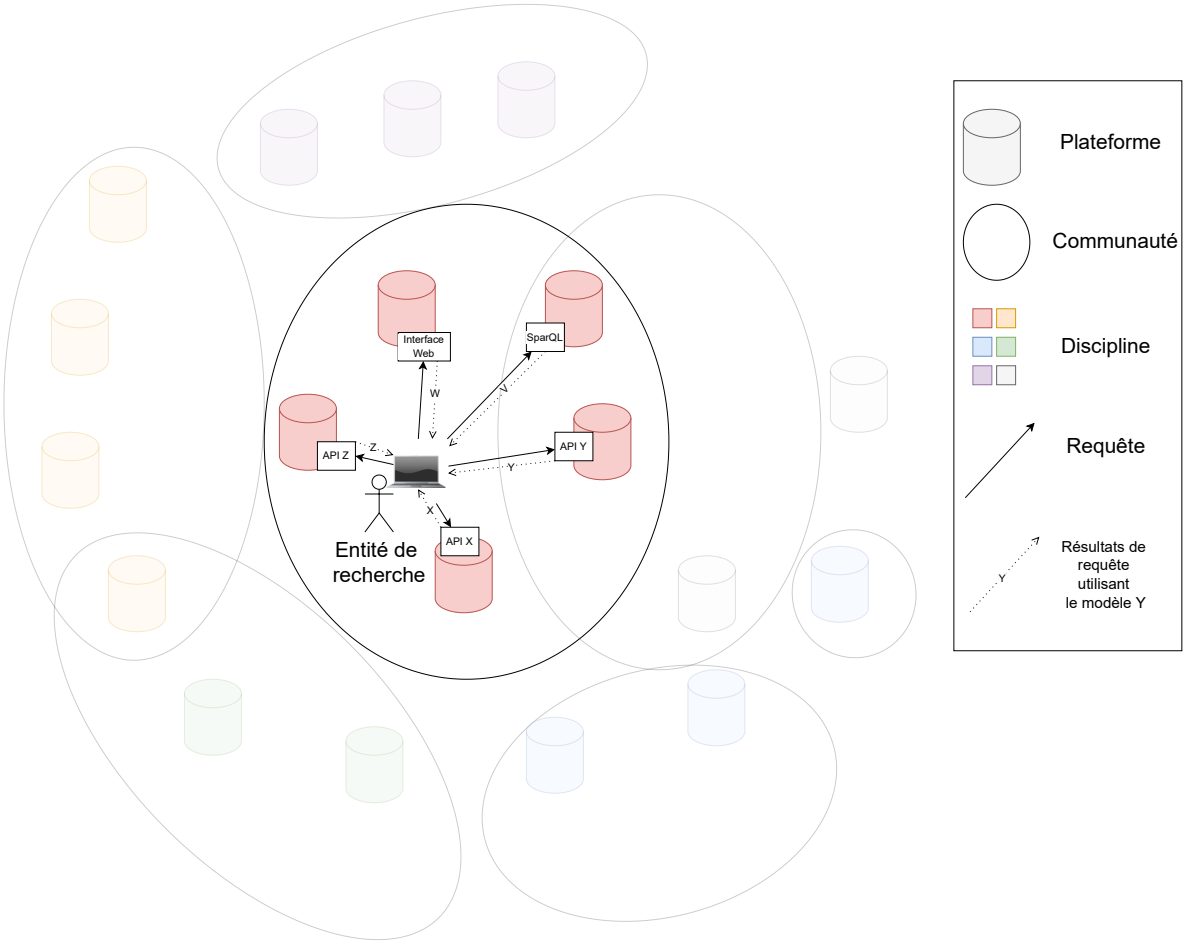

FIGURE 4.1 – Recherche d'information intracommunautaire par le chercheur P

Pour permettre cette recherche intracommunautaire de données de recherche, nous proposons d'étendre le concept de Lac de Données au contexte de la Science Ouverte. Habituellement, un lac de données est une solution de Big Data Analytics définie comme un emplacement unique et centralisé permettant de stocker et de gérer à bas coût de grands volumes de différents types de données, dans leur format natif à mesure qu'elles arrivent (Ravat and Zhao (2019a)). Le concept de lac de données nécessite plusieurs adaptations pour permettre un partage de métadonnées intracommunautaire et devenir un Lac de Données de la Science Ouverte (LDSO) : (1) la gestion de nouvelles sources de données avec les sources externes, (2) la modification de la gestion des métadonnées dans la zone de gouvernance et (3) une intégration au cœur de l'architecture fonctionnelle de mécanismes de sécurité des données pour répondre au risque posé par l'ouverture des données (Peisert et al. (2017)).

Pour proposer cette solution de partage intracommunautaire de données, nous proposons le concept de LDSO (Section 4.1). Tout d'abord, nous définissons l'architecture fonctionnelle du LDSO (Section 4.2). Nous précisons les nouvelles sources de données (Section 4.2.1) et la nouvelle gestion des métadonnées basées sur une gestion multi-modèles (Section 4.2.2) d'un LDSO. Nous présentons aussi les mécanismes de contrôle

d'accès et d'authentification ajoutés via différents composants nécessaires à un partage de métadonnées intracommunautaire dans la Science Ouverte. Enfin, nous proposons une architecture technique permettant l'implantation du LDSO (Section 4.3). Nous terminerons en montrant comment le LDSO répond à la problématique d'interopérabilité des plateformes de gestion de données dans la Science Ouverte.

4.1 Définition

Le lac de données est habituellement défini comme "une solution de big data analytics permettant différents types d'utilisateurs (data scientist, analystes de données, professionnels de la business intelligence, etc...) de [réaliser des ingestions de données brutes hétérogènes, préparer les données pour les nouveaux besoins des entreprises, réaliser différentes analyses sur ces données et gouverner les données pour permettre leur qualité, leur sécurité et leur cycle de vie]". (Zhao (2021))

Le Lac de Données répond à la diversité de besoins d'analyses et la variété de types de données nécessaires aux différentes communautés de la recherche scientifique. Cette solution permet un accès unifié aux données, demandé par des chercheurs. Cependant, plusieurs verrous de la Science Ouverte ne sont pas considérés dans la définition standard du lac de données :

- le problème de passage à l'échelle qui est posé par le volume existant dans la Science Ouverte rend impossible une intégration physique de toutes les données, approche prise dans la définition standard du Lac de Données.
- la grande variété de modèles de métadonnées ainsi que l'existence d'instances de métadonnées dans les données sources nécessitent une modification du processus de gouvernance.
- la sécurité des lacs de données n'est pas pensée pour permettre son ouverture et l'ouverture de ses données.

Dans le chapitre 2, nous avons étudié deux solutions de lacs de données avec CoreKG (Beheshti et al. (2018)) et la proposition de Castro et al. (2022) qui propose une architecture logicielle de référence pour la Science Ouverte basée sur un lac de données. Ces deux solutions se basent sur l'utilisation d'un modèle unique pour les métadonnées ou sur l'utilisation d'un ensemble de standards interconnectés. Dans le chapitre 2, nous avons observé que la standardisation des modèles de métadonnées ne permet pas de résoudre le problème de variété des API, ce qui empêche ces solutions de répondre à la problématique d'interopérabilité des PDRO dans la Science Ouverte. De plus, ces solutions ne prennent pas en compte la contrainte du volume trop important de la Science Ouverte et donc ne passent pas à l'échelle. Ces solutions ne répondent donc pas aux besoins de la Science Ouverte.

Ainsi, nous proposons d'étendre le concept de lac de données pour proposer le Lac de Données de la Science Ouverte (LDSO).

Définition - Lac de Données de la Science Ouverte (LDSO) : *Un Lac de Données de la Science Ouverte est un espace dédié aux différentes analyses (entraînement de modèles d'intelligence artificielle, jointure de jeux de données pour réaliser des analyses croisées, recherche d'articles scientifiques pour la mise en place de revues systématiques, etc...) de données de recherche des chercheurs offrant les fonctionnalités suivantes :*

- *ingérer physiquement ou virtuellement des données de recherche pour offrir une recherche unifiée de données de recherche des PDRO d'une communauté scientifique*

en répondant à la problématique de l'interopérabilité des PDRO ;
— *transformer ces données de recherche à la demande, pour répondre à de nouveaux besoins de recherche en manipulant ces données ;*
— *permettre la gestion d'utilisateurs internes et externes à l'entité de recherche en enrichissant les différents processus de gestion des données dans l'architecture avec des processus de sécurité (identification, authentification et contrôle d'accès) ;*
— *ouvrir les données à la demande des entités de recherche.*

Le LDSO doit être un lac de données sécurisé, ouvert et profitant de l'existence de plateformes externes grâce à une interopérabilité avec ces PDRO.

4.2 Architecture fonctionnelle

Dans cette section, nous définissons l'architecture fonctionnelle du LDSO, en mettant en avant les spécificités du contexte de Science Ouverte. La Figure 4.2 illustre l'architecture fonctionnelle du LDSO.

4.2.1 Les données sources

Le LDSO doit gérer des **données locales**, provenant des travaux de recherche d'entités de recherches internes, mais aussi des **données externes**, non stockées localement et provenant de plateformes de gestion de données existantes dans la Science Ouverte.

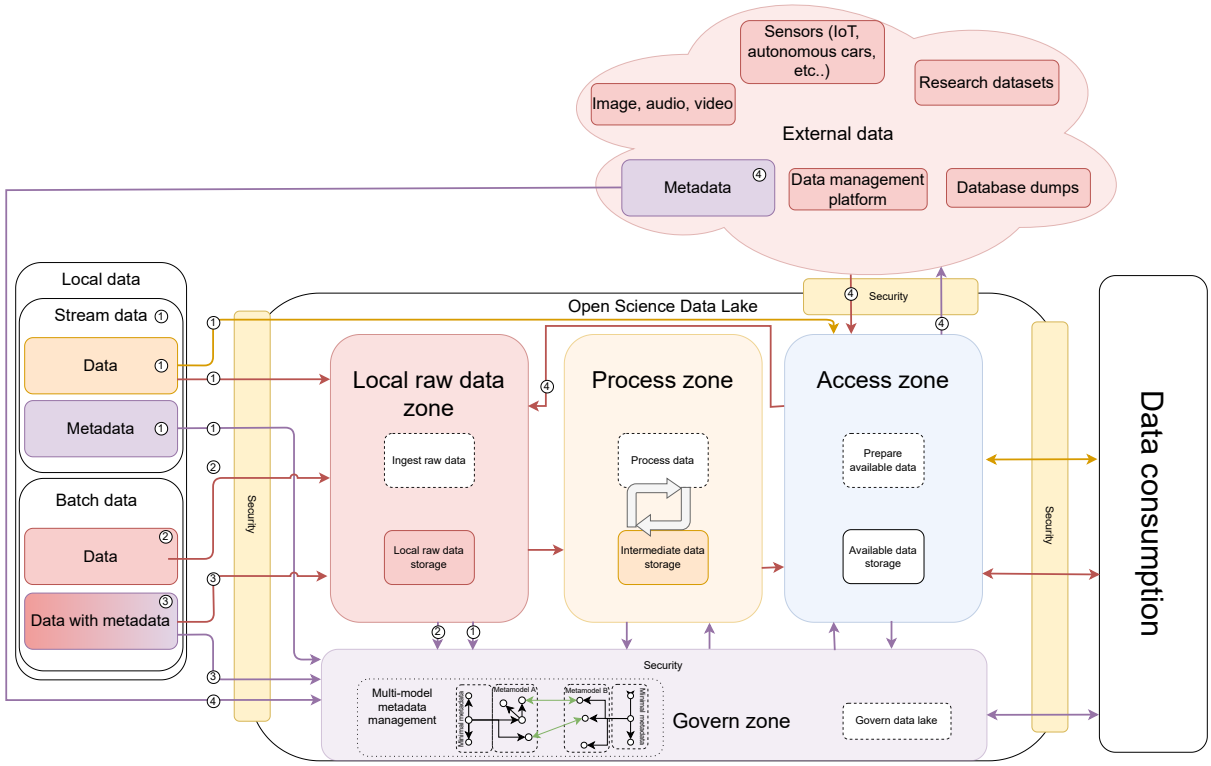

FIGURE 4.2 – L'architecture fonctionnelle de Lac de Données de la Science Ouverte

4.2.1.1 Les données locales

Nous proposons de décomposer les données locales en trois catégories (cf Figure 4.2) :

— (1) les données par flux, potentiellement avec des contraintes temporelles : Le flux est initialisé grâce aux métadonnées stockées dans la zone de gestion de métadonnées. Une fois le flux mis en place, les données peuvent ainsi arriver en flux et être directement consommées dans la zone d'accès. Elles peuvent aussi être ingérées dans la zone de données brutes pour être analysées sur du long terme. Par exemple, supposons l'entité de recherche de l'IRIT. Cette entité de recherche gère un projet de recherche basé sur des données de capteurs dans le cadre de l'opération neOCampus (Gleizes et al. (2018)). Ces données de capteurs arrivent à intervalles réguliers, en mesurant des grandeurs physiques (température, humidité, CO2, etc...) dans les différents bâtiments de l'université. Ces données sont collectées à partir d'un environnement contrôlé et nourissent plusieurs projets de recherches (par exemple : l'étude des dynamiques de flux de personnes en fonction des mesures effectuées). Ces données sont des données par flux.

— (2) les données par lot sans métadonnées : ces données sont reçues et ingérées dans la zone d'ingestion. Les métadonnées sont générées au fur et à mesure du passage des données dans les différentes zones fonctionnelles (Ravat and Zhao (2019a)), permettant la surveillance du cycle de vie des données. Prenons le cas d'une entité de recherche souhaitant réaliser une publication scientifique sur un problème de recherche. Cette publication scientifique se base sur un ensemble de données que l'entité de recherche possède localement. Cette entité de recherche dépose ses données dans la zone d'ingestion permettant la préparation et la consommation de ces données dans les différentes zones fonctionnelles. Ces données sont des données par lot et les métadonnées sont générées au fur et à mesure, permettant de suivre les différentes analyses effectuées sur ces données.

Contrairement aux lacs de données classiques, le LDSO gère une troisième catégorie de type de données avec les données ayant déjà été gérées par des plateformes de la Science Ouverte et possédant déjà une vie avant son intégration dans le LDSO :

— (3) les données par lot avec métadonnées : Ces données sont reçues et ingérées dans la zone de données brutes. Les métadonnées qui accompagnent ces données sont intégrées dans la zone de gouvernance. Ces métadonnées sont enrichies tout au long du cycle de vie des données, permettant de conserver l'historique des opérations effectuées dans les autres PDRO tout en suivant le cycle de vie au sein du LDSO. Les modèles de métadonnées utilisés sont variés et une réponse doit être proposée pour répondre à cette variété pour assurer l'interopérabilité. Prenons la réalisation d'une étude systématique sur les travaux de définition de l'interopérabilité (Maciel et al. (2024)). Cette revue systématique nécessite la récupération de différentes publications scientifiques provenant de différentes PDRO. Chaque publication scientifique étudiée est ingérée dans la zone d'ingestion et les métadonnées associées dans la zone de gouvernance. Les opérations effectuées sur ces publications (par exemple : exploration ou text mining) sont ajoutées aux instances de métadonnées insérées avec les publications. Ces données sont des données par lot avec métadonnées.

Ce nouveau type de données engendre des modifications dans le processus d'ingestion dans la zone de données brutes et dans la gestion des métadonnées de la zone de gouvernance pour permettre de gérer la variété de modèles de métadonnées, que nous détaillons dans la suite.

4.2.1.2 Les données externes

L'ingestion de la totalité des données présentes sur les plateformes de gestion de données de la Science Ouverte pose un problème de passage à l'échelle. Pour traiter ce problème, nous ajoutons une intégration virtuelle des données des autres PDRO. Cette intégration virtuelle permet d'ajouter des données à la liste des données connues du PDRO sans avoir à les copier localement grâce à l'ingestion de leurs métadonnées. Nous définissons donc un nouveau type de données à gérer par le LDSO avec les données externes :

— (4) les métadonnées seules : les métadonnées associées à des données de recherche stockées sur d'autres PDRO sont intégrées directement dans la zone de gouvernance. Ces données sont intégrées virtuellement dans le système de gestion de métadonnées du LDSO, permettant d'indexer une grande quantité de données à faible coût dans le catalogue de données du LDSO. Ces données doivent être téléchargées depuis la plateforme les hébergeant avant leur utilisation. Prenons l'exemple de la PDRO OpenDataSoft [1]. Le catalogue de données de cette PDRO est téléchargé et ingéré dans le LDSO. Cette ingestion de métadonnées permet, au moment de la recherche dans LDSO, de réaliser une recherche sur les données locales du LDSO mais aussi sur la PDRO OpenDataSoft. Le résultat de cette recherche de données locales et externes est retourné à l'utilisateur de façon transparente et permet d'enrichir les données connues dans le LDSO à faible coût.

Ce nouveau type de données permet une intégration virtuelle de données et nécessite des modifications dans les zones de traitement et d'accès, pour gérer le stockage externe des données dans le processus de consommation des données.

4.2.2 La zone de gouvernance

La zone de gouvernance est "en charge d'assurer la sécurité des données, la qualité des données, le cycle de vie des données, l'accès aux données et de la gestion des métadonnées" Ravat and Zhao (2019b); Zhao (2021). Pour s'adapter à ces nouvelles données en entrée du lac de données, notamment à la variété des leur modèles de métadonnées, nous proposons de modifier la gestion des métadonnées, la sécurité des données et les processus d'accès aux données pour intégrer les nouveaux profils de données mais aussi pour répondre aux besoins de sécurité liés à l'ouverture des données.

4.2.2.1 Gestion multi-modèles des métadonnées

Le système de gestion de métadonnées dans la zone de gouvernance d'un lac de données est défini comme "un système qui est basé sur des métadonnées standards et qui permettent de générer et maintenir les métadonnées de façon automatique et d'explorer ces métadonnées avec un système ergonomique" (Zhao (2021)). Cette définition est inadaptée au contexte de la Science Ouverte pour deux raisons :

— la standardisation des métadonnées ne permet pas de répondre à l'hétérogénéité des modèles de métadonnées ;
— l'intégration virtuelle de données n'est pas incluse.

1. `https://www.opendatasoft.com/fr/`

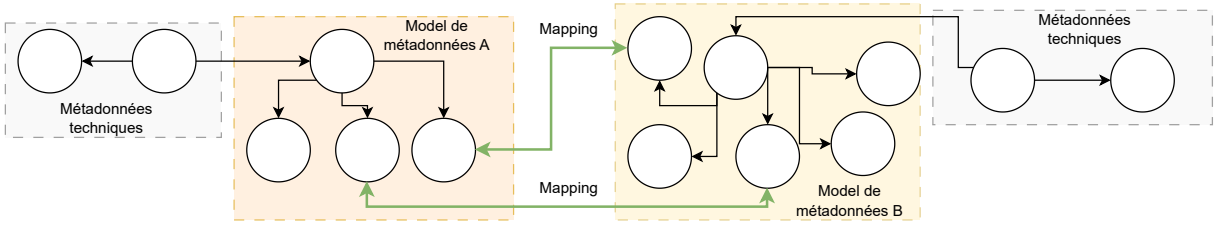

FIGURE 4.3 – Gestion multi-modèles dans le LDSO

Définition : Système de gestion de métadonnées du LDSO *Le système de gestion de métadonnées du LDSO est un système dans la zone de gouvernance permettant l'intégration de métadonnées basée ingestion sans modification des métadonnées et sur une gestion multi-modèles pour permettre l'utilisation de ces métadonnées. L'objectif du système de gestion de métadonnées du LDSO est soit de générer et maintenir automatiquement les métadonnées sur les données de recherche, soit d'ingérer et enrichir automatiquement des métadonnées existantes. De plus, ce système doit fournir un accès transparent aux métadonnées et une exploration transparente de ces métadonnées.*

Pour réaliser l'intégration de modèles hétérogènes et fournir un accès transparent aux métadonnées de différentes PDRO, la gestion des métadonnées dans le LDSO est basée sur une gestion multi-modèles avec l'utilisation de mappings entre ces modèles.

Cette gestion multi-modèles se base sur l'utilisation sans modification des modèles ou des instances de métadonnées utilisés pour les données ingérées. Les métadonnées sont ingérées sans modification (cf partie orange ou jaune dans la Figure 4.3). Un mécanisme d'interopération manuel par mise en place de passerelles entre les concepts des différents modèles de métadonnées est ensuite appliqué, avec un mapping de ces concepts pour définir les concepts égaux (cf les flèches vertes dans la Figure 4.3). Ces mappings sont réalisés à tout moment par les utilisateurs ou les gestionnaires de plateformes. Cette interopération permet l'utilisation transparente des métadonnées.

Pour les données locales ne possédant pas de métadonnées (les données par lots sans métadonnées et les données par flux), ces métadonnées doivent être générées, avec le même processus qu'un lac de données classique. Pour générer les métadonnées, un modèle est choisi par l'utilisateur ou peut être défini par défaut. Puis les métadonnées sont générées en instanciant ce modèle et enrichies au fur et à mesure du cycle de vie des données dans le LDSO.

Ensuite, ces métadonnées sont enrichies avec un ensemble de métadonnées techniques uniquement nécessaires au bon fonctionnement des services internes aux LDSO (cf partie grise dans la Figure 4.3). Ces métadonnées techniques sont générées automatiquement et servent aux différents services du LDSO pour fonctionner. Ces métadonnées techniques contiennent au minimum :

- un identifiant unique au sein du LDSO ;
- un nom de fichier ;
- une localisation, afin de retrouver les données avant leur consommation mais aussi de distinguer les données locales des données externes ;
- un propriétaire, nécessaire à la mise en place de mécanismes de contrôle d'accès aux ressources du lac de données ;
- une date d'insertion dans le lac de données.

Le modèle pour ces métadonnées techniques est commun à tous les jeux de données

du LDSO. L'interopération est assurée grâce à la standardisation sur cet ensemble de métadonnées uniquement utiles aux services internes du LDSO.

Exemple : Nous illustrons la gestion multi-modèles à l'aide de deux modèles de métadonnées. Tout d'abord, le modèle utilisé dans le système DAMMS (Zhao (2021)) est retenu comme modèle par défaut lors de l'ingestion de données locales sans métadonnées. Ce modèle est adapté aux lacs de données et permet une gestion de l'ensemble des étapes du cycle de vie des données. Ensuite, nous prenons un modèle hypothétique pour la gestion des traitements effectués : le modèle Algorithm Metadata Vocabulary (AMV) défini par Dutta and Patel (2021). Cet exemple illustre comment l'interopération des modèles permet une recherche des données de recherche en se basant sur les informations des analyses effectuées sur différentes plateformes de gestion de données. Par souci de simplicité, nous n'avons sélectionné pour cet exemple qu'une sous-partie des modèles en lien avec les algorithmes et leurs implantations.

La Figure 4.4 illustre l'intégration du modèle utilisé dans le DAMMS dans le LDSO. Le modèle du DAMMS est utilisé sans modification (cf. les classes orange dans la Figure 4.4). Les métadonnées techniques nécessaires aux services internes du LDSO sont ajoutées à ces métadonnées (cf. la classe grise dans la Figure 4.4). Dans cet exemple, ce modèle est utilisé pour l'ensemble des données de recherche locales sans métadonnées.

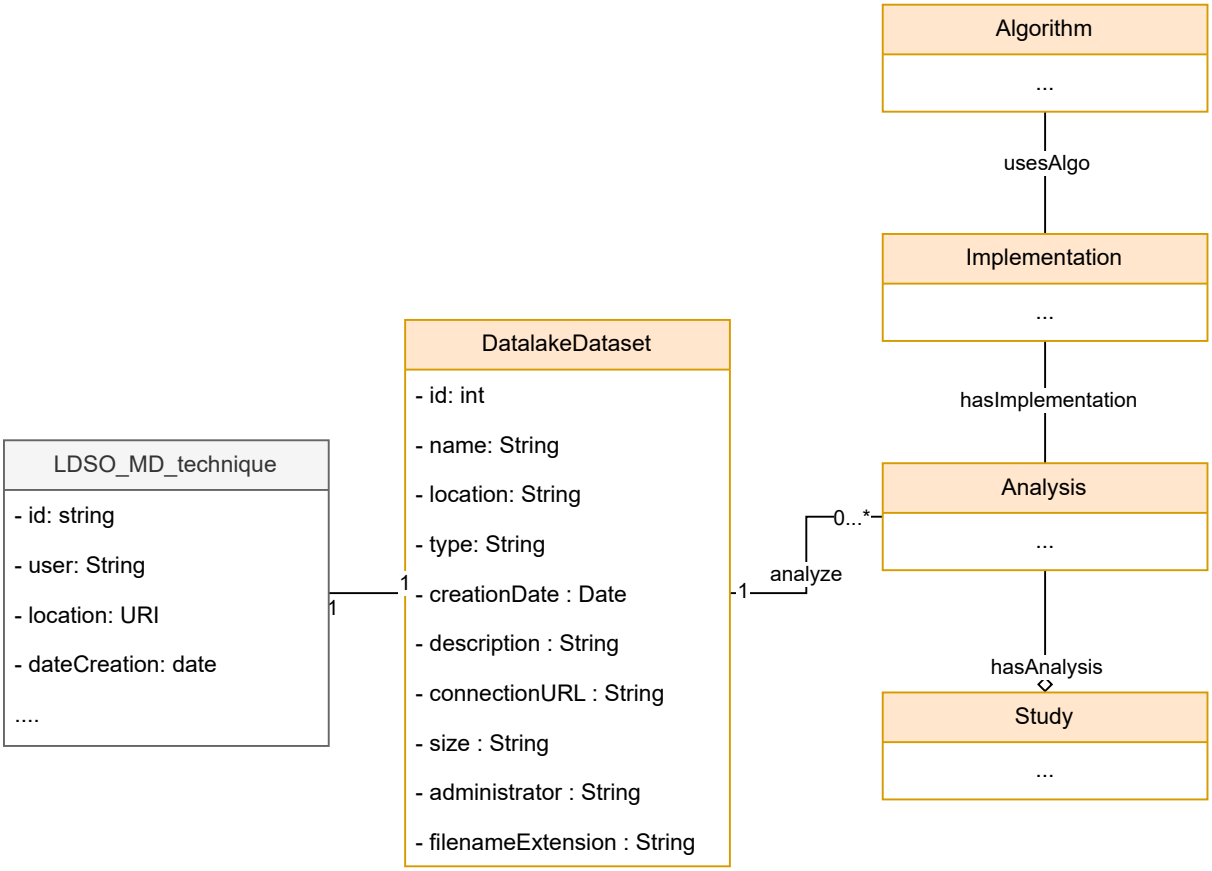

FIGURE 4.4 – Intégration de métadonnées suivant le modèle du DAMMS dans le LDSO

Supposons maintenant qu'un utilisateur du LDSO souhaite intégrer des données de recherche faisant partie d'une autre PDRO, qui utilise le modèle AMV. Les métadonnées

sont ingérées sans modification du modèle afin de conserver les informations descriptives sur ces données de recherche (cf partie orange dans la Figure 4.5). Comme pour le modèle du système DAMMS, les métadonnées sont enrichies par l'ensemble minimal de métadonnées techniques (cf. partie grise sur la Figure 4.5).

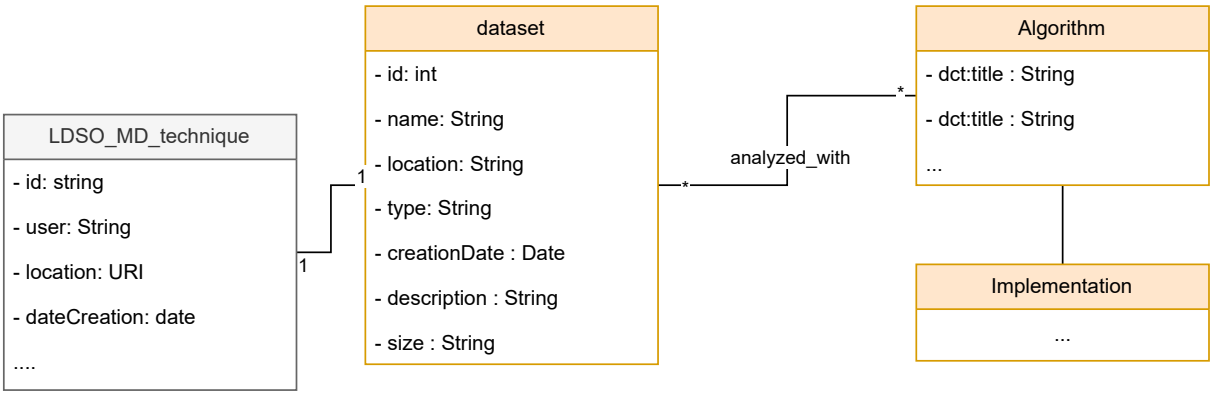

FIGURE 4.5 – Intégration de métadonnées suivant un modèle incluant le modèle AMV dans le LDSO

Les mappings sont effectués sur les entités liées aux algorithmes utilisés ainsi que sur leur implantation (cf. les flèches vertes dans la Figure 4.6). Grâce à ces mappings, la re-

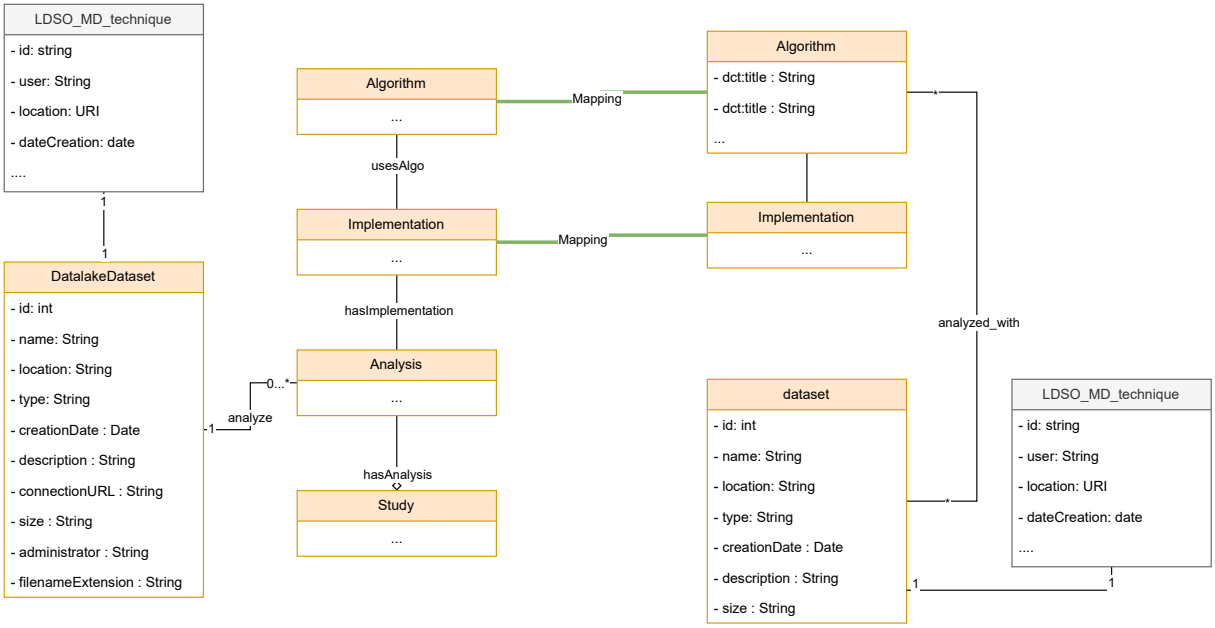

FIGURE 4.6 – Interopération des modèles de métadonnées intégrés dans le LDSO

cherche de données de recherche traitées par un certain algorithme permet de trouver des données locales et externes suivant les conditions d'application d'un algorithme spécifique et de son implantation. Ces données peuvent potentiellement être croisées, car répondant aux mêmes conditions. **Une recherche d'information intracommunautaire, portant sur des données internes et externes, est possible** grâce à ce mécanisme.

4.2.2.2 Sécurisation des accès

L'ouverture des données implique l'ouverture du processus à des utilisateurs externes à l'entité de recherche ciblé par le déploiement du LDSO. Ces utilisateurs sont des entités de recherche externes ne faisant pas partie de l'entité de recherche locale et donc ne possédant pas d'identité au sein du système de gestion des utilisateurs du LDSO. L'ouverture des données implique aussi une ouverture du réseau sur lequel le LDSO est déployé. Cette ouverture empêche la mise en place de mécanismes pour limiter les accès au réseau local permettant une sécurisation indirecte de la plateforme. Pour traiter ce problème, nous incorporons au coeur même du LDSO **des mécanismes de contrôle d'accès et d'authentification**. Ces mécanismes sont effectués en bordure du LDSO, sur l'ensemble des points d'entrée du LDSO (cf parties jaunes dans la Figure 4.2).

Ces mécanismes de gestion des accès reposent sur une politique d'accès basée sur les groupes et un système d'authentification permettant de conserver les informations sur les utilisateurs et les groupes associés. Nous pouvons mettre en place un premier mécanisme de contrôle d'accès aux ressources (données et services proposés) dans le LDSO. Ce sont les identifiants de ces utilisateurs qui sont conservés dans les métadonnées techniques. Cette approche est classique en gestion des accès.

Les données de recherche sont considérées par défaut comme fermées et uniquement accessibles au propriétaire de ces données, pour suivre le comportement par défaut en cas d'absence de licence du code logiciel. L'information sur l'ouverture est à trouver dans le modèle de métadonnées utilisé pour décrire ces données de recherche.

Du point de vue des utilisateurs, chaque demande de ressource suit ce cheminement, décrit dans un schéma BPMN dans la Figure 4.7 :

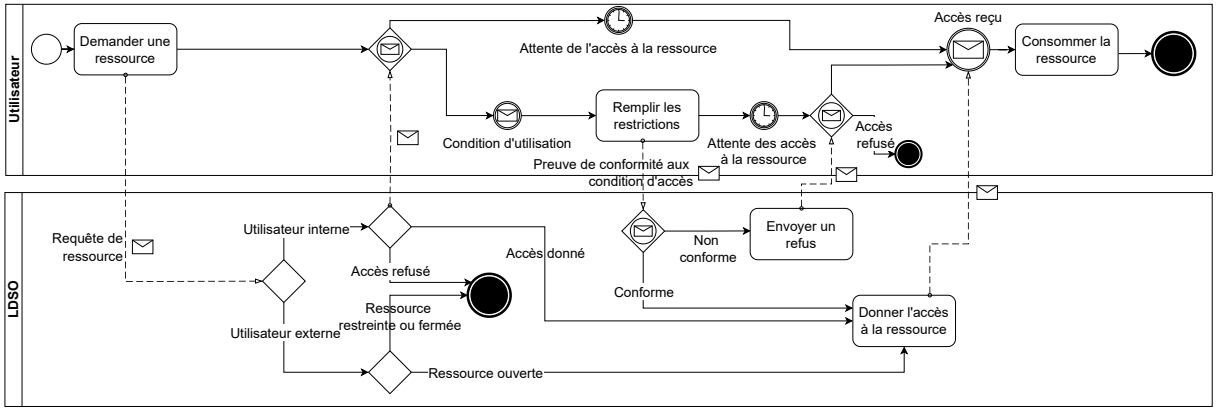

FIGURE 4.7 – Processus de demande d'accès à une ressource par un utilisateur

— (1) La requête d'une ressource est reçue par le LDSO. Pour les utilisateurs externes, seules les ressources ouvertes (données ouvertes, services ouverts) sont accessibles. L'utilisateur interne (inscrit sur la plateforme et connu du système d'authentification) peut accéder à l'ensemble des ressources, à condition qu'il remplisse les conditions associées à la ressource (par exemple : appartenir à un projet spécifique).
— (2) Dans le cas des ressources restreintes, une demande de conformité aux ressources est envoyée à l'utilisateur. Par exemple, les informations sur les utilisations envisagées de la ressource ou une signature de la charte d'utilisation peuvent être demandées à l'utilisateur.

— (3) Une fois les conditions remplies par l'utilisateur ou dans le cas d'une ressource ouverte, l'accès à la ressource est accordé.

Si une étape n'est pas respectée, l'accès à la ressource est refusé.

Du point de vue du LDSO, la demande d'accès se fait avec deux groupes de services :

— les services de gestion des accès (composés de l'outil d'accès aux services et du système d'authentification) ;

— le système de gestion de métadonnées.

Cette demande d'accès est illustrée dans le schéma BPMN dans la Figure 4.8. Le système de gestion des accès récupère les informations liées aux utilisateurs (le(s) groupe(s), le(s) rôle(s), le(s) projet(s) de l'utilisateur) auprès du système d'authentification.

En parallèle, le système de contrôle d'accès récupère dans le système de gestion de données les métadonnées d'accès sur une ressource si elles existent dans les modèles de métadonnées (licence, utilisateurs propriétaires, niveau d'ouverture, etc...). En cas d'absence de ces informations, l'accès est limité aux propriétaires de la donnée. Le croisement de ces différentes informations permet d'observer, pour une ressource donnée, les quatre scénarios suivants de terminaison de ce processus :

— l'utilisateur est un utilisateur non enregistré (un utilisateur externe) et n'accède à la ressource que si la ressource est ouverte ;

— l'utilisateur est un utilisateur enregistré (un utilisateur interne) et il possède les droits suffisants pour accéder à la ressource ;

— l'utilisateur est un utilisateur enregistré et la ressource est fermée, l'utilisateur n'a accès à la ressource que s'il en est l'un des propriétaires.

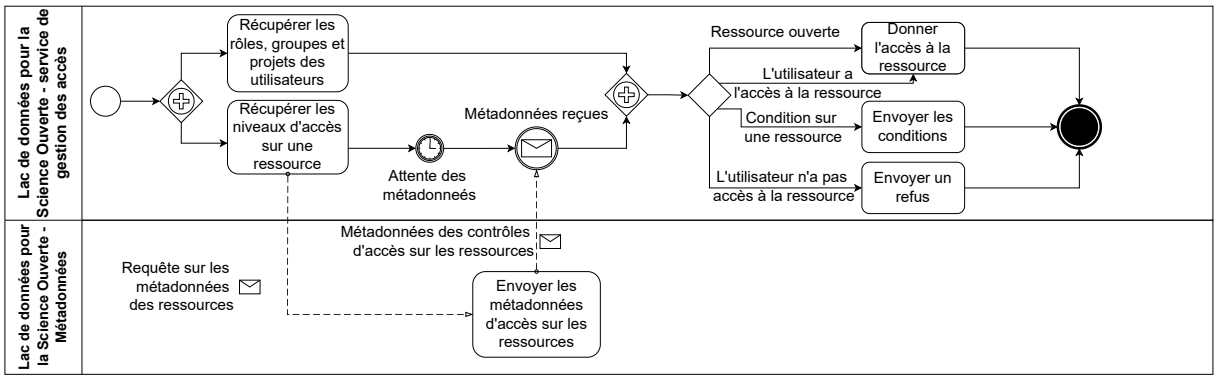

FIGURE 4.8 – Processus interne de gestion des accès aux ressources dans le LDSO

Ces mécanismes de sécurité assurent **l'intégrité et la confidentialité, nécessaires à des collaborations public / privé ou la mise en place de brevet** avec ces données. Ces mécanismes permettent de **mettre en place une ouverture des données à la demande des propriétaires**.

4.2.3 Zone d'ingestion

Les modifications apportées par les nouveaux types de données nécessitent une extension des processus d'ingestion (Zhao et al. (2021)). La Figure 4.9 est un graphique BPMN illustrant les différents processus d'ingestion du LDSO. L'ingestion des données déjà présentes dans les lacs de données classiques est conservée :

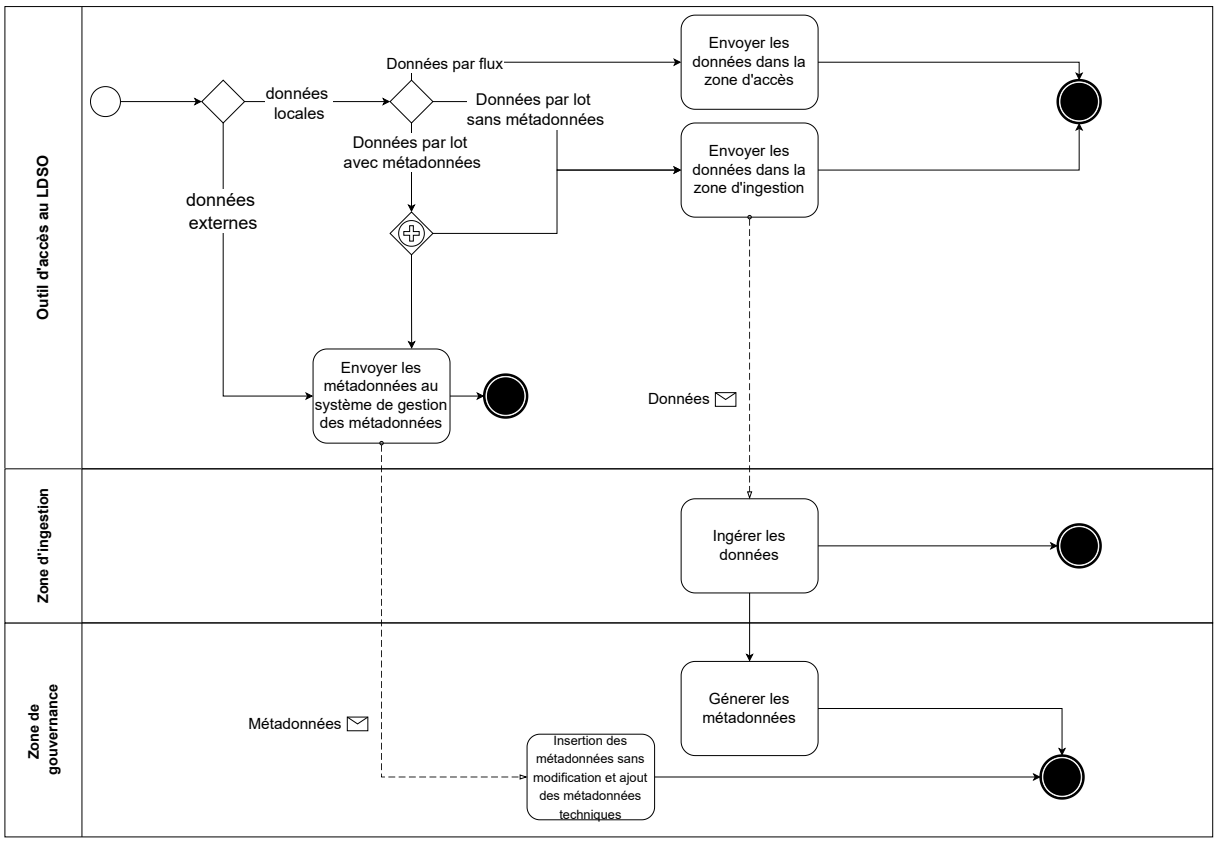

FIGURE 4.9 – Processus d'ingestion du LDSO selon les différents types de données sources

— L'ingestion des données locales par flux (cf les flèches numérotées (1) dans la Figure 4.2) est identique au processus classique des lacs de données (Zhao et al. (2021)). Les métadonnées permettant l'instanciation du flux de réception des données sont ingérées dans le système de gestion de métadonnées. Le flux de données est ensuite reçu directement dans la zone d'accès pour être consommé, permettant une réduction des délais de consommation. En parallèle, ces données sont stockées dans la zone d'ingestion pour permettre un stockage pérenne de ces données.
— L'ingestion des données locales par lot sans métadonnées (cf les flèches numérotées (2) dans la Figure 4.2) est identique au processus classique des lacs de données (Zhao (2021); Ravat and Zhao (2019b)). Les données sont stockées dans la zone d'ingestion sans modification et les métadonnées sont ensuite générées pour suivre le cycle de vie de cette donnée.

Un nouveau processus de gestion de données locales est ajouté au LDSO pour gérer les données de recherche provenant d'autres PDRO :

— Les données locales avec des métadonnées (cf les flèches numérotées (3) dans la Figure 4.2) sont stockées dans la zone d'ingestion sans modification. Les métadonnées sont insérées dans la zone de gouvernance sans modification. Des métadonnées techniques sont ajoutées à ces métadonnées. Ensuite, ces métadonnées servent de base pour suivre le cycle de vie de ces données dans le LDSO.

Des mappings sur ces modèles doivent être présents au moment de la consommation de données pour permettre une utilisation transparente des métadonnées.

Pour permettre l'intégration virtuelle de données externes, nous avons un dernier pro-

cessus d'ingestion ne s'appliquant que sur les métadonnées :

— Les données externes (cf les flèches numérotées (4) dans la Figure 4.2) ne sont composées que de métadonnées. Ces métadonnées sont insérées dans la zone de gouvernance.

4.2.4 Zone de traitement et zone d'accès

La zone de traitement et la zone d'accès sont enrichies pour ajouter les mécanismes de sécurité et de gestion de métadonnées multi-modèles au sein de leur processus de fonctionnement.

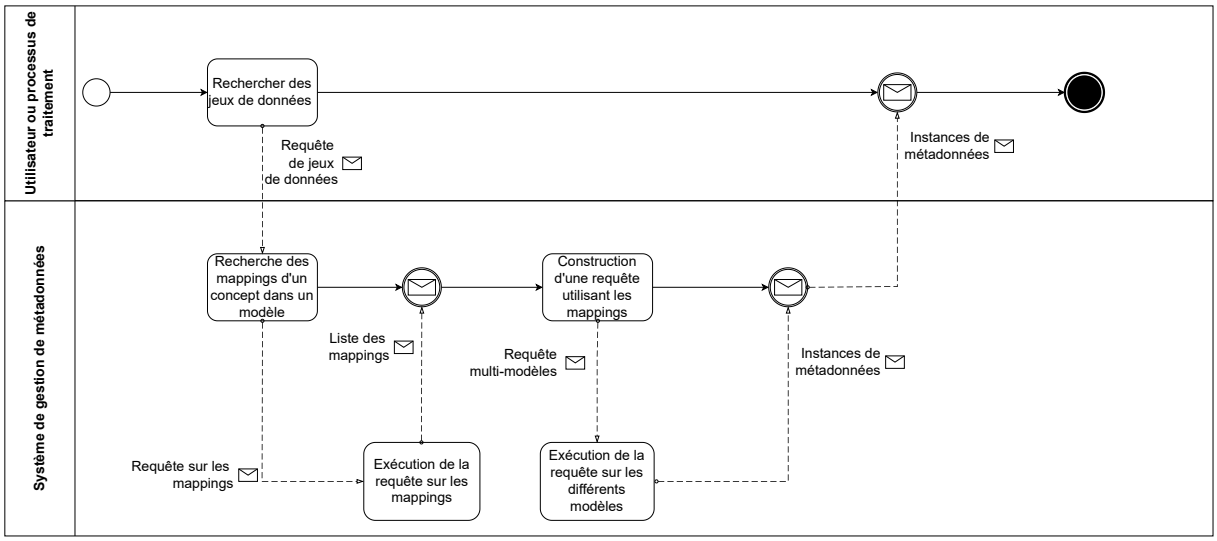

FIGURE 4.10 – Processus de recherche de données de recherche dans le LDSO avec une gestion multi-modèles

L'utilisation aux métadonnées ne repose plus sur un modèle unique. L'accès aux concepts des métadonnées doit permettre un accès transparent aux métadonnées, en utilisant les mappings entre les concepts. Le mécanisme de recherche de métadonnées est modifié en ajoutant une recherche des mappings associés puis l'utilisation de ces mappings pour générer les requêtes associées. Ce mécanisme suit le processus illustré dans la Figure 4.10. Lors de la recherche de jeux de données par un utilisateur ou un processus de traitement, le système de gestion de métadonnées reçoit une requête. Avant l'éxecution de cette requête, une recherche de mappings pouvant exister avec le modèle utilisé pour cette requête est réalisée. Une requête est construite, basée sur les différents mappings trouvés, et l'ensemble des métadonnées est retourné à l'utilisateur. Cette recherche transparente permet de trouver des données quel que soit le modèle utilisé, à partir du moment où des mappings sont réalisés.

Les données externes nécessitent une étape supplémentaire de téléchargement des données avant leur consommation. Une fois téléchargées, ces données externes sont consommées dans la zone d'accès et ingérées dans la zone d'ingestion pour permettre un stockage pérenne des données. Les métadonnées associées sont modifiées, notamment pour indiquer la nouvelle localisation des données en local. Le processus de récupération de données est décrit dans la Figure 4.11.

Lors de la demande de récupération de données par un utilisateur ou un processus de traitement, les métadonnées techniques des données demandées sont récupérées. Les données locales sont téléchargées depuis la zone d'ingestion et utilisées comme le processus classique. Pour les données externes, les données sont téléchargées depuis la PDRO qui les gère. Ensuite, ces données sont ingérées dans la zone d'ingestion.

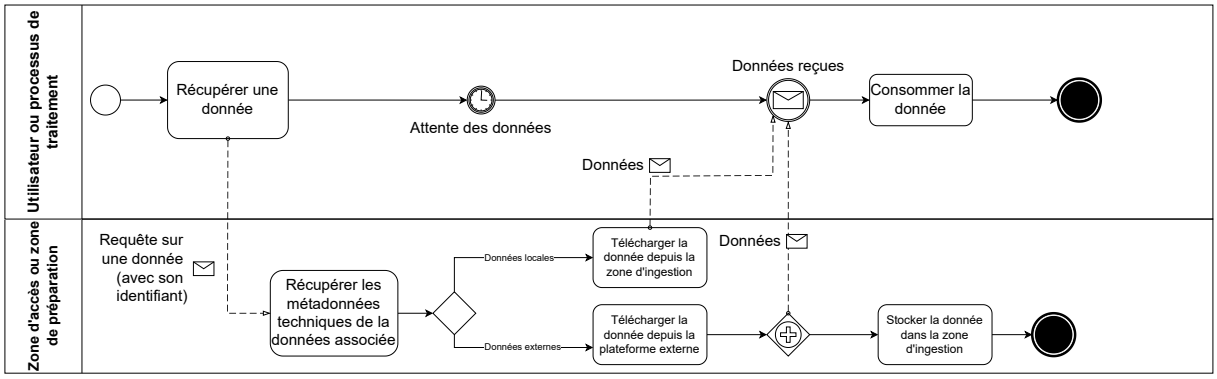

FIGURE 4.11 – Processus de gestion des données locales et externes

4.3 Architecture technique

Maintenant que nous avons présenté l'architecture fonctionnelle, l'implantation de ces mécanismes nécessite de préciser certaines caractéristiques. La gestion multi-modèles nécessite que le système de gestion de données puisse gérer plusieurs modèles de métadonnées et les mappings associés. Les services de la zone de préparation et d'accès doivent être suffisamment flexibles pour permettre une modification des processus classiques du lac de données.

Nous proposons une architecture technique visant l'implantation du LDSO et répondant à ces besoins. La totalité des outils déployés dans cette architecture technique sont conteneurisés avec la technologie Docker. Les technologies de conteneurisation, comme celles de virtualisation, permettent un déploiement automatique et simplifié de solutions logicielles grâce à leur portabilité multiplateformes [2].

En dehors des besoins fonctionnels auxquels les outils doivent répondre, nous appliquons deux critères à l'ensemble des outils :

— Les outils sélectionnés sont open-sources.
— Les outils sélectionnés possèdent une communauté importante et ayant un développement long terme.

Nous avons principalement sélectionné des outils provenant de deux organisations (Openstack et Apache) qui développent et maintiennent des outils open-sources depuis plusieurs dizaines d'années. Le développement de ces outils est soutenu par de nombreuses grandes entreprises du monde de l'informatique.

Dans cette section, nous utilisons la Figure 4.14 pour illustrer tous les outils utilisés pour chaque zone. Nous définissons les outils de la zone de gouvernance (cf. Section 4.3.1), puis les outils de la zone d'ingestion (cf. Section 4.3.2), puis les outils de la zone de traitement (cf. Section 4.3.3), puis les outils de la zone d'accès (cf. Section 4.3.4).

2. `www.docker.com/resources/what-container`

4.3.1 Zone de gouvernance

La zone de gouvernance est composée de deux types d'outils :

— Le système de gestion de métadonnées est réalisé avec MongoDB, une base de données NoSQL orientée document . Cette base de données est dite "sans schéma" [3] ("schemaless") permettant une intégration de documents au format JSON avec des champs différents dans une même collection. Grâce à cette caractéristique, nous pouvons ingérer **dans une même collection, des métadonnées avec des modèles différents.**

— Le système de gestion des authentifications avec Openstack Keystone. Ce système de gestion des accès permet une personnalisation du choix de backend utilisé permettant une adaptation à un large panel de technologies d'authentifications. Cet outil permet **la définition de rôles, de groupes et de projets nécessaire à notre gestion des accès basée sur des rôles**. Pour la mise en place des accès, nous avons mis en place une API RESTful basée sur la bibliothèque Flask. Ces deux outils permettent de gérer l'ensemble des accès en ajoutant les processus de contrôle d'accès à toutes les ressources du LDSO.

MongoDB permet la gestion "schemaless" permettant d'avoir des documents avec plusieurs modèles de métadonnées différents au sein d'une même collection. L'implantation de cette gestion multi-modèles est donc possible avec un faible coût d'utilisation, sans avoir à gérer différentes bases de données utilisant un modèle différent et sans modification à mettre en place des documents en entrée. La base de données est composée de trois collections distinctes :

— la collection "metadonnées" permettant de stocker les instances de métadonnées, avec leurs différents modèles. Chaque instance est enrichie avec les métadonnées techniques uniquement en ajoutant des champs aux documents stockés ;

— la collection "mapping" permettant de stocker les mappings entre les concepts ;

— la collection "modele" permettant de stocker l'ensemble des modèles de métadonnées utilisables par les utilisateurs.

Pour gérer ces mappings, nous ajoutons un champ "modele" aux métadonnées techniques, qui est l'identifiant unique des modèles dans la collection "modele", permettant d'indiquer quel est le modèle utilisé.

Les fonctions de recherche standardisées sont définies dans Apache Airflow et permettent avant toute recherche de métadonnées sur un concept de vérifier si une correspondance est trouvée avec le concept d'un autre modèle. À l'issue de cette recherche, une requête est définie sur l'ensemble des champs conservant le même concept. Ce mécanisme est décrit dans la Figure 4.12).

Nous illustrons le résultat de ce mécanisme dans la Figure 4.13, avec un exemple de requête permettant de trouver l'ensemble des données ayant utilisé un algorithme nommé "Algorithme X" avec le langage de requêtage de MongoDB.

4.3.2 La zone d'ingestion

Les lacs de données sont souvent associés à la technologie Hadoop pour la gestion des données. Cette technologie permet de réaliser des traitements hautement parallèles basés sur le concept de Map/Reduce. Cependant, cette technologie possède une grande latence, incompatible avec la gestion de flux de données, et encore plus avec les flux de données

3. `https://www.mongodb.com/resources/basics/unstructured-data/schemaless`

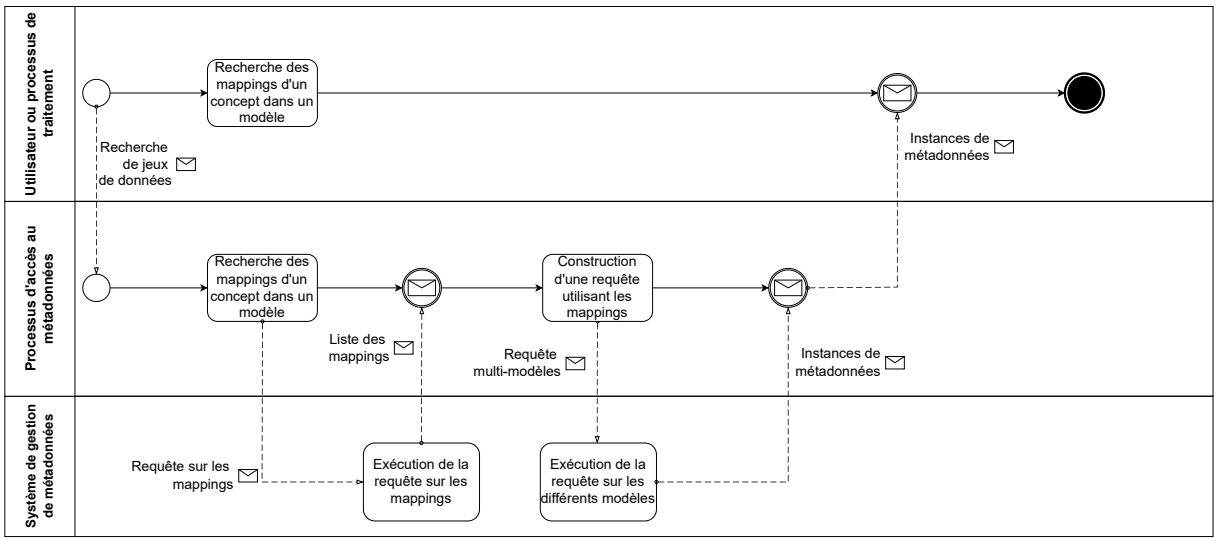

FIGURE 4.12 – Processus technique de recherche de données de recherche avec une gestion multi-modèles

```
"db.metadonnee.find({
    $or :[
        {
        "dataset.algorithm.dct:title" : { $in : "Algorithme X"}
        },
        {
        "DatalakeDataset.analysis.implementation.algorithm.name :
            {$in : "Algorithme X"}
        }
    ]

})
```

FIGURE 4.13 – Exemple de requête multi-modèles dans la base de données MongoDB

ayant des contraintes temporelles. De plus, cette technologie se base sur le système de fichier HDFS qui crée des blocs de stockage des données de 128 Mo par défaut. Cette approche ne s'adapte pas à la gestion des données de l'IoT pouvant peser quelques octets mais étant en très grand nombre. Ces raisons nous ont poussés à concevoir une solution sans l'environnement Hadoop. La zone d'ingestion est basée sur un stockage orienté objet, avec Openstack Swift. Cet outil permet de gérer un grand volume de données à faible coût et gère les données sous forme d'objets stockés sans modification. Cet outil permet donc de gérer un grand volume, une grande variété et propose des capacités de passage à l'échelle suffisantes pour répondre aux besoins de la zone d'ingestion quel que soit le type de données à ingérer.

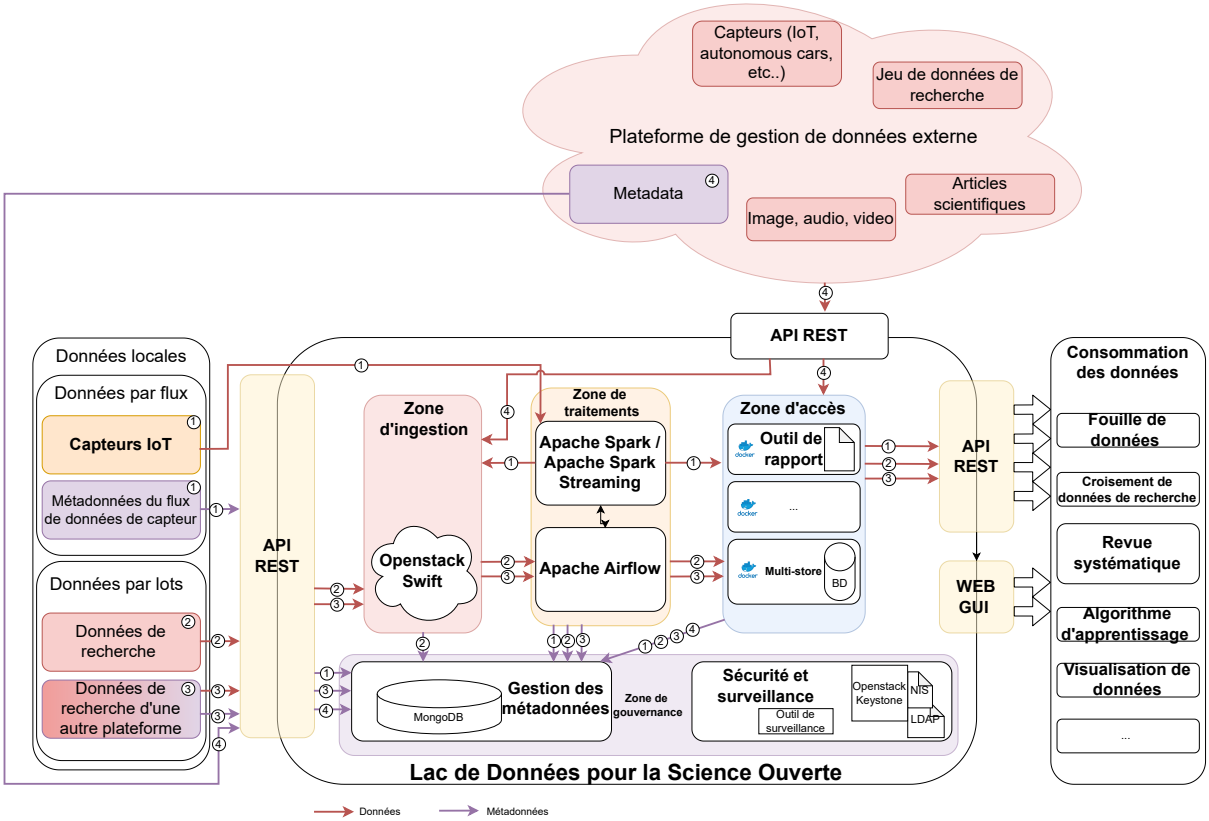

FIGURE 4.14 – Architecture technique de lac de données

4.3.3 La zone de traitement

La zone de traitement est composée de deux outils (cf Zone de traitement dans la zone 4.14) :

- **Apache Airflow** est un orchestrateur de pipeline de traitements. Cet outil sert à la mise en place de l'ensemble des processus de traitement, d'analyse et de mise à jour des métadonnées. Il assure toute l'orchestration des différents services entre eux, mais permet aussi aux utilisateurs de définir des pipelines de traitements à appliquer en fonction d'un type de données, d'une source ou autre. Cet outil permet d'**la réutilisation et l'automatisation des traitements effectués sur les données**. Il permet aussi la mise en place de traitements récurrents, pouvant être utilisés pour **l'échange d'information intracommunautaire semi-automatique avec les plateformes**.
- un cluster de calcul **Apache Spark** permettant le traitement de données en grand volume, grâce à ses capacités de déploiement de traitement hautement parallèle. Le choix de cet outil permet en plus de gérer les données de flux grâce à **Apache Spark Streaming**. Les flux sont reçus directement par Apache Spark Streaming et peuvent être traités et consommés dans un délai faible.

La Figure 4.15 illustre l'organisation des traitements dans Airflow, selon les utilisateurs, les projets et les types de données à traiter. Dans cet exemple, la préparation des données peut être réalisée selon le projet (DataNoos, IDEAS ou neocampus) et selon le type de données (archive, CSV, données batchs, données de stream, BSON). De plus, des traitements par

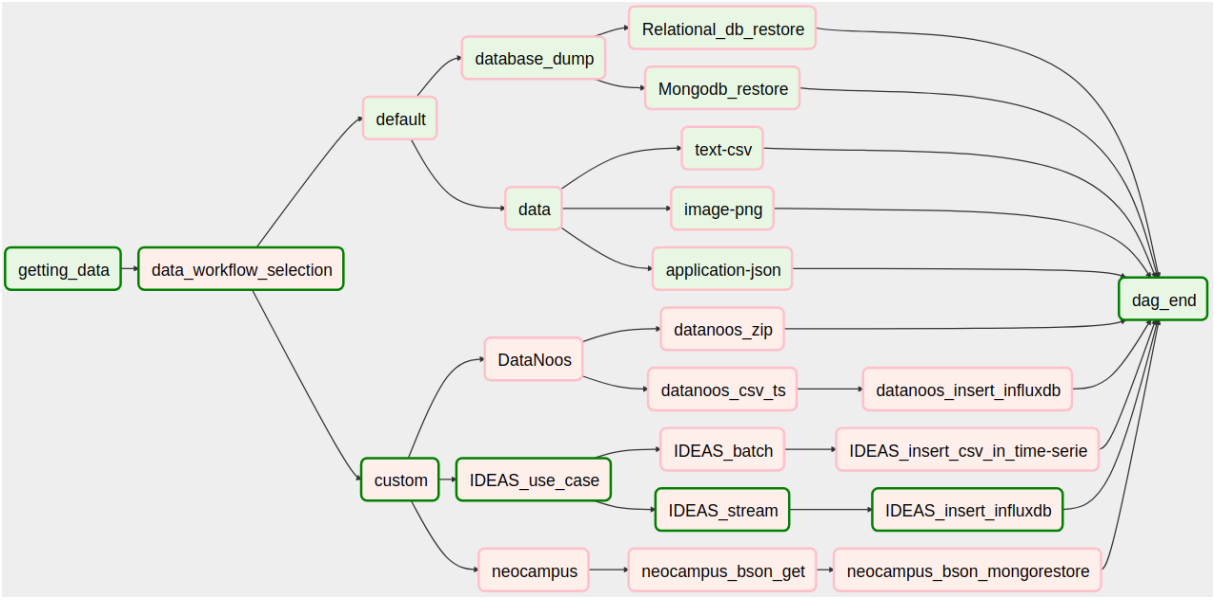

FIGURE 4.15 – Exemple de pipeline de traitement : gestion de deux projets différents (neOCampus, DataNoos et un exemple pour la conférence IDEAS)

défaut sont définis et proposés aux utilisateurs.

4.3.4 La zone d'accès

Cette zone d'accès est conçue comme une zone d'accueil de conteneurs Docker. Ainsi, n'importe quel outil de traitement, d'analyse ou de consommation de données est déployable pour les utilisateurs. De cette manière, nous répondons à la **grande variété de besoins de traitement présents dans le contexte de la Science Ouverte**. Par exemple, cette zone peut accueillir des moteurs de base de données conteneurisés, des outils de dashboarding, des outils d'analyses comme des notebooks Python, indépendants pour chaque projet. Cette séparation permet d'assurer sécurité et confidentialité. De plus, les technologies de conteneurisation offrent une facilité et une automatisation du déploiement permettant de modifier facilement les outils déployés.

4.4 LDSO et interopérabilité

Pour permettre un partage de données de recherche intracommunautaire, nous devons répondre à la problématique de l'interopérabilité des PDRO. Cette interopérabilité nécessite de répondre aux deux variétés observées : (1) la variété d'API de communication et (2) la variété de modèles de métadonnées. Dans cette section, nous décrivons comment le LDSO permet de répondre à ces deux variétés.

4.4.1 Hétérogénéité des API : approche hybride d'interopération du LDSO

Pour l'hétérogénéité des API de communication des plateformes de gestion de données ouvertes, nous avons choisi une approche hybride d'interopération, qui mélange mécanisme

d'interopérabilité par standardisation et par mise en place de passerelles.

Dans le LDSO, nous avons choisi de mettre en place une API RESTful pour la gestion des accès et des communications. En effet, les API REST sont utilisées par une majorité de PDRO dans la recherche (cf Chapitre 2 et section 3.3.2.3).

Une adaptation de l'API REST est nécessaire pour intégrer la variété de chemins d'accès (par exemple, l'URL pour requêter les métadonnées d'une plateforme) et de formats des messages HTTP des autres API REST (couches L5 à L7). Cependant, la réduction du nombre de plateformes à intégrer par l'aspect intracommunautaire permet une mise en place manuelle de ces chemins d'accès et des formats de messages.

Pour les autres plateformes ne disposant pas d'API REST par défaut, nous avons proposé une API REST dans l'architecture technique qui est basée sur la bibliothèque Flask. Cette bibliothèque Python permet le développement d'API RESTful simples et modifiables. Ainsi, des connecteurs à d'autres types d'API (par exemple des connecteurs vers des API SparQL) sont disponibles dans les bibliothèques communautaires Python. Ces bibliothèques logicielles fournissent des mécanismes d'interopérabilité utilisant des passerelles entre les différentes API.

4.4.2 Hétérogénéité des modèles de métadonnées : implantation de la gestion multi-modèles

L'approche adoptée pour gérer l'hétérogénéité des modèles de métadonnées est une approche hybride d'interopération, qui combine standardisation et utilisation de passerelles. L'ensemble des modèles est conservé dans le système de gestion de métadonnées du LDSO. Ces modèles sont ensuite interopérés grâce à des passerelles utilisant les mappings. Nous avons ajouté à chaque instance de métadonnées des métadonnées techniques pour les besoins techniques des services. Ces métadonnées techniques se basent sur une standardisation de la modélisation utilisée.

Notre approche permet de proposer un accès aux métadonnées transparent mais nécessite une modification des processus de gestion des métadonnées dans le LDSO, comme observé dans les sections précédentes.

4.5 Conclusion

La recherche de données intracommunautaires permet aux entités de recherche de trouver des données de recherche sur des sujets en lien avec la discipline de cette entité. Ainsi, les entités de recherche peuvent réutiliser des données de recherche, potentiellement traitées et nettoyées pour répondre à de nouvelles questions de recherche avec par exemple (1) l'entraînement de modèles d'intelligence artificielle qui nécessitent un grand volume de données, (2) la jointure de jeux de données de recherche pour réaliser des analyses croisées ou (3) la recherche d'articles scientifiques pour effectuer une revue de question systématique.

La multiplication des PDRO et des modèles de métadonnées utilisés complexifie la recherche de données de recherche. Sans partage intracommunautaire de métadonnées, des données de recherche peuvent ne pas être trouvées par les entités de recherche qui en ont besoin. De plus, le nombre de PDRO à prendre en main peut être important et empêcher l'exploration exhaustive de ces données.

Pour permettre le partage intracommunautaire de métadonnées nécessaire pour une recherche unifiée de données, nous avons proposé une extension du concept de lac de

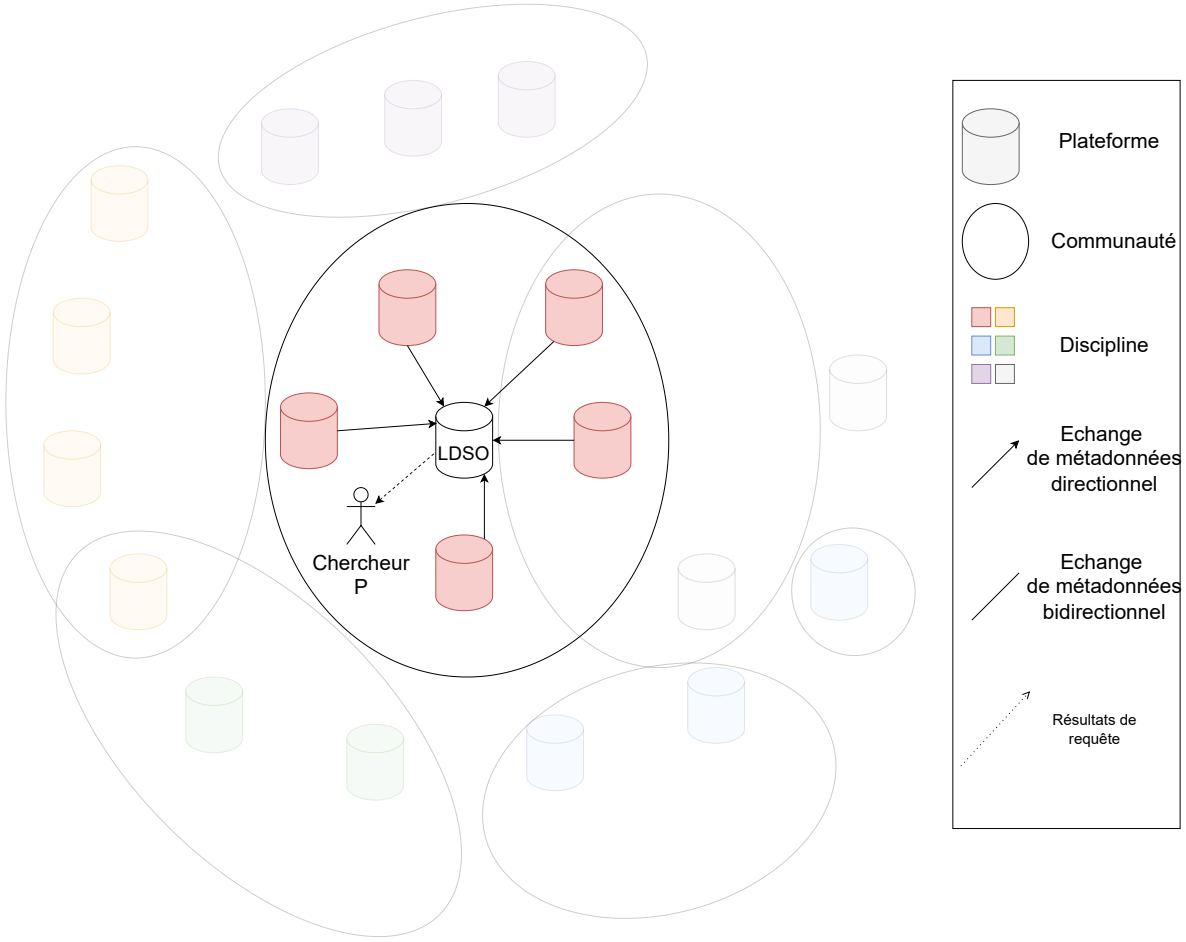

FIGURE 4.16 – Impact du LDSO sur la Science Ouverte : mise en place d'un échange d'information intracommunautaire

données avec le Lac de Données de la Science Ouverte (LDSO). Nous avons proposé une définition du LDSO et une architecture fonctionnelle du LDSO. Pour s'adapter aux spécificités de la Science Ouverte et de la recherche intracommunautaire de données, nous avons intégré au LDSO les éléments suivants :

— une gestion multi-modèles de métadonnées, basée sur des mappings entre les modèles, fournissant un accès unifié aux métadonnées ;
— des mécanismes de contrôle d'accès et d'authentification permettant la gestion du nouveau profil de données engendré par l'ouverture des données (les utilisateurs externes) et offrant aux propriétaires de données un contrôle sur l'ouverture des données ;
— une intégration virtuelle de données permettant de gérer le trop grand volume dans la Science Ouverte.

Nous avons ensuite proposé une architecture technique, en définissant l'ensemble des outils utilisés, qui sont tous "open-source". Pour l'implantation de l'interopérabilité des PDRO, le LDSO se base sur deux choix :

— Une interopérabilité hybride pour répondre à la variété des API, utilisant d'une

part la standardisation d'une grande partie de PDRO qui ont choisi une API REST et d'autre part des passerelles avec les autres types d'API, permises par les API REST ;
— Une interopérabilité s'appuyant sur des passerelles pour répondre à la variété des modèles de métadonnées, utilisant des mappings entre les concepts des modèles de métadonnées.

En reprenant l'exemple de l'entité de recherche ER, la proposition du LDSO permet une réduction du coût de la recherche intracommunautaire. La figure 4.16 illustre la communauté d'ER après l'implantation d'un LDSO avec les catalogues de données des PDRO téléchargées et ingérées dans le LDSO. Le LDSO offre à ER une recherche unifiée dans les PDRO de sa communauté, grâce à l'intégration virtuelle des données. Il n'a plus qu'à apprendre à utiliser un seul modèle et une seule PDRO pour accéder à l'ensemble des données des PDRO de sa communauté. Le coût de la recherche est réduit pour ER.

Nous validons la capacité d'échange intracommunautaire, ainsi que la réduction du coût de la recherche de données, dans le chapitre 6. Nous proposons une expérimentation basée sur une preuve de concept et une expérience utilisateur afin de valider la capacité de recherche transparente intracommunautaire du LDSO et les réductions de coûts apportées par le LDSO.

Dans notre exemple, nous avons utilisé un mécanisme de partage de métadonnées manuel, avec l'ingestion des catalogues de métadonnées gérées par des entités de recherche interne au LDSO autre que ER. Ce partage de métadonnées manuel pose un problème de passage à l'échelle qui n'était pas présent dans le contexte réduit de l'intracommunautarité. Mais lors de la recherche de données interdisciplinaire et intercommunautaire, le passage à l'échelle devient un frein majeur. Pour compléter le LDSO sur le mécanisme de partage de métadonnées dans ce contexte, notre solution vise l'échange de données interdisciplinaire et intercommunautaire grâce à un réseau d'interconnexion de PDRO dans le chapitre 5. Le LDSO a fait l'objet d'une publication scientifique dans la conférence internationale ADBIS 2023 (Dang et al. (2023b)).

Chapitre 5

Recherche interdisciplinaire et intercommunautaire : le Réseau de Données de la Science Ouverte

L'interdisciplinarité est perçue comme essentielle à la recherche et de plus en plus présente dans les questions de recherche abordées par les chercheurs (Corbett et al. (2013); Ramachandran et al. (2021)). Cette recherche de données interdisciplinaire permet :

— Un enrichissement des réponses apportées à une question de recherche
— Un enrichissement des questions de recherche explorées

Ces apports proviennent des nouveaux outils reposant sur l'exploration de données provenant de différentes disciplines et/ou domaines.

Pour illustrer ce constat, nous reprenons l'entité de recherche ER (cf. Figure 5.1). ER a répondu à sa question de recherche initiale (cf. Chapitre 4) grâce aux recherches de données intracommunautaires permises par le LDSO. ER souhaite maintenant explorer de nouvelles questions de recherche. Pour ce faire, ER veut explorer les données de recherche d'autres communautés ou disciplines, qui lui permettrait d'enrichir ses approches potentielles sur des questions de recherche. Mais ER ne connaît pas les PDRO des autres communautés ou disciplines (cf. points d'interrogation dans la figure 5.1). ER doit donc mettre en place une recherche préalable de ces PDRO. Deux situations sont possibles dans ce cas :

— ER arrive à trouver une plateforme avec des données ouvertes d'une autre communauté ou discipline. Il doit alors accéder à cette PDRO - à condition que l'accès soit ouvert -, apprendre à utiliser les outils proposés par cette PDRO, avec des modélisations spécifiques, comprendre les types de données, télécharger les données puis les ingérer dans le LDSO pour leur utilisation. L'utilité de ces données peut être difficile à évaluer pour ER, qui n'a potentiellement pas les compétences pour comprendre rapidement les formats et/ou le contenu de ces données de recherche. De plus, ce processus doit être répété jusqu'à ce qu'ER trouve des données de recherche qu'il comprenne et qui lui soient utiles. Une telle activité peut donc lui être coûteuse en temps.
— ER n'arrive pas à trouver de plateforme avec des données ouvertes d'autres disciplines, il ne peut donc pas élargir ses activités de recherche dans une perspective interdisciplinaire.

En l'état, la recherche interdisciplinaire d'ER soulève les mêmes problématiques que la recherche intracommunautaire sans LDSO, avec un trop grand nombre de plateformes

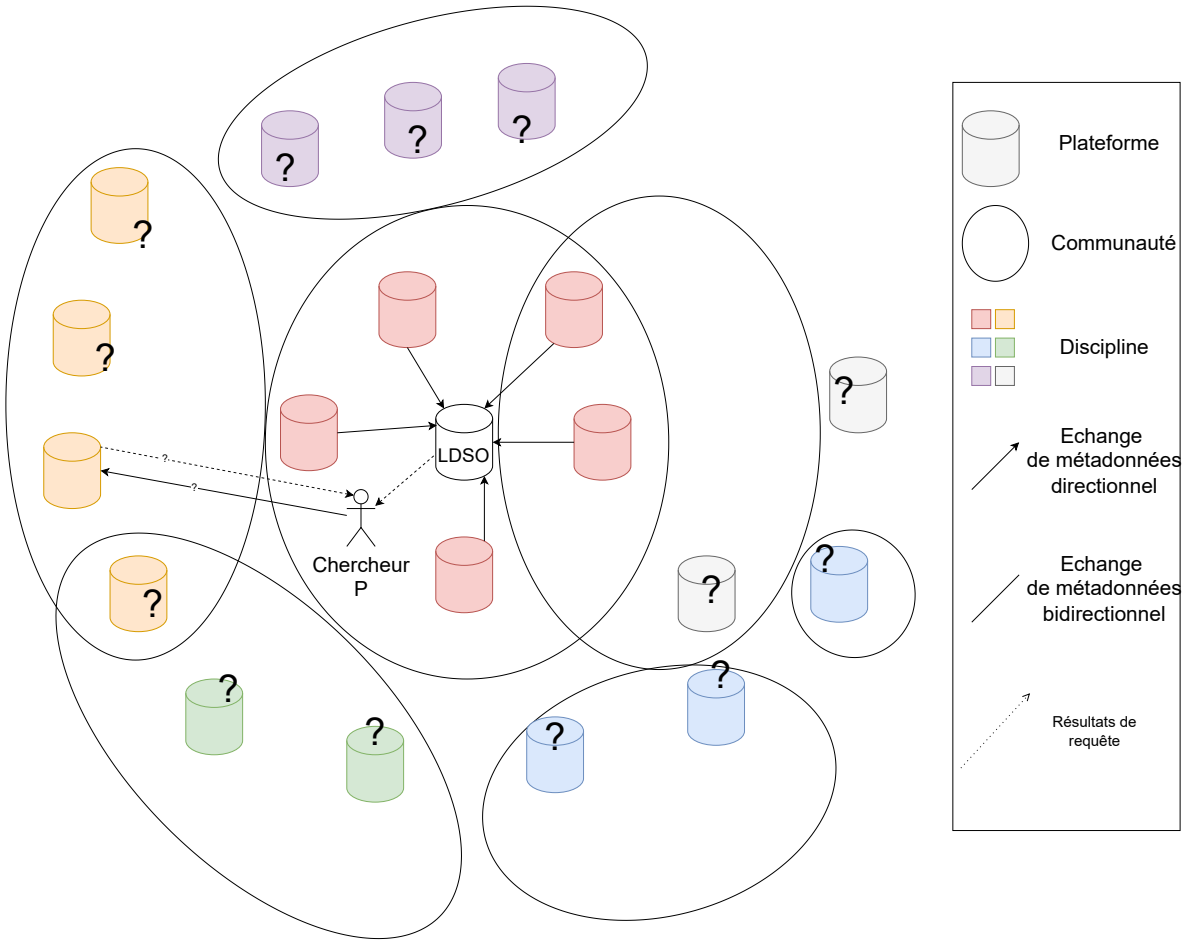

FIGURE 5.1 – Exemple d'ER : problème de la recherche de données interdisciplinaire et intercommunautaire

à connaître et à savoir utiliser. Une problématique supplémentaire se pose pour ER : la connaissance et le coût d'utilisation de nouvelles plateformes spécifiques à des disciplines ou communautés qu'il ne connaît pas.

Nous avons observé dans le chapitre 3 que l'échange de métadonnées dans la Science Ouverte, nécessaire à une recherche d'information unifiée à l'ensemble des PDRO, n'est pas implantée à l'heure actuelle. Les chercheurs ont décrit plusieurs problématiques à la mise en place de cette recherche de données :

— La recherche de données est trop coûteuse pour les chercheurs par un manque de ressources (Van Panhuis et al. (2014); Top et al. (2022); Rainey et al. (2023); Sadeh et al. (2023); European Commission and Directorate-General for Research and Innovation et al. (2021))
— Le contrôle sur l'ouverture des données est primordial (Top et al. (2022); Sadeh et al. (2023); Kathawalla et al. (2021); Rainey et al. (2023))

Des solutions présentées dans le chapitre 2 ont pour objectif de réaliser des PDRO pour l'ensemble des entités de recherche quelle que soit la discipline, telle l'EOSC. Cependant, ces solutions se basent sur :

— une standardisation des modèles de métadonnées, qui ne répond pas à la variété de modèles de métadonnées (cf. Chapitre 2)
— une centralisation qui possède des limites de passage à l'échelle (Van Steen and Tanenbaum (2023))

Le LDSO (cf. chapitre 4) a permis de proposer une solution de recherche de données unifiée dans la Science Ouverte (cf. Fig 5.1). Dans le LDSO, le partage de métadonnées est :

— unidirectionnel, des PDRO vers le LDSO, n'offrant pas la possibilité aux utilisateurs des PDRO de profiter de cet échange d'information (une interopérabilité non réciproque) ;
— intracommunautaire, ne permettant pas d'enrichir les connaissances de cette communauté avec des données provenant d'autres communautés ou d'autres disciplines non incluses dans cette communauté à cause du problème de passage à l'échelle.

Cependant, la recherche de données interdisciplinaires demande de prendre en compte cette problématique de passage à l'échelle. Pour permettre une recherche interdisciplinaire, nous souhaitons donc proposer une solution complète qui assure une recherche unifiée de données de recherche interdisciplinaire et intercommunautaire répondant à la problématique du passage à l'échelle.

Pour cela, nous proposons le concept de *Réseau de Données de la Science Ouverte* (RDSO) ; le RDSO est un réseau décentralisé, distribué et fédéré d'interconnexion de PDRO. Le RDSO permet de **rechercher des données unifiées** dans l'ensemble des PDRO, en se basant sur les métadonnées de celles-ci. Cette recherche de données est basée sur un mécanisme d'échange de métadonnées entre plateformes. Ce réseau présente l'avantage d'être décentralisé. Cette décentralisation permet aux PDRO d'être autonomes et de répartir les charges à l'ensemble des acteurs de la recherche scientifique grâce à un effort communautaire et fédéré.

Dans la section 5.1, nous présentons notre proposition de RDSO au travers de ses composants (cf. Section 5.1.1) et ses fonctionnalités (cf. Section 5.1.2). Dans la section 5.2, nous présentons les approches prises pour implanter l'interopérabilité des PDRO avec le RDSO, avec une réponse à la variété des API de communication (cf. Section 5.2.1) et une réponse à la variété des modèles de métadonnées (cf. Section 5.2.2). Dans la section 5.3, nous comparons notre solution à une autre architecture décentralisée visant une consommation de données de plusieurs domaines d'une entreprise, au travers du prisme des données ("data mesh"). Cette comparaison est nécessaire pour positionner notre proposition par rapport aux propositions proches.

5.1 Architecture d'un RDSO

La recherche de données interdisciplinaires nécessite la mise en place d'un échange de métadonnées entre les PDRO pour que la recherche de données soit unifiée à toutes ces PDRO. Pour permettre ce partage de métadonnées et cette recherche unifiée de données, nous proposons le Réseau de Données de la Science Ouverte (RDSO). Le RDSO est un réseau décentralisé, distribué et fédéré d'interconnexions de PDRO. Il a pour objectif de favoriser une exploration unifiée, interdisciplinaire et intercommunautaire de données de recherche sur l'ensemble des PDRO. Cette solution se base sur trois composants : (1) les plateformes (ou PDRO), (2) le module du RDSO et (3) un registre distribué. Ces composants permettent la définition de fonctionnalités avec (i) un passage à l'échelle, (ii) une sécurité des échanges de données et (iii) une mise en place d'un mécanisme d'échange d'in-

formation semi-automatique entre les plateformes. Dans la section 5.1.1, nous détaillons les trois composants du RDSO. Dans la section 5.1.2, nous détaillons les fonctionnalités du RDSO.

5.1.1 Les composants du RDSO

Le RDSO est une architecture de réseau permettant l'interopérabilité des PDRO grâce à un mécanisme d'échange de métadonnées qui facilite une recherche de données inter-communautaire et interdisciplinaire.

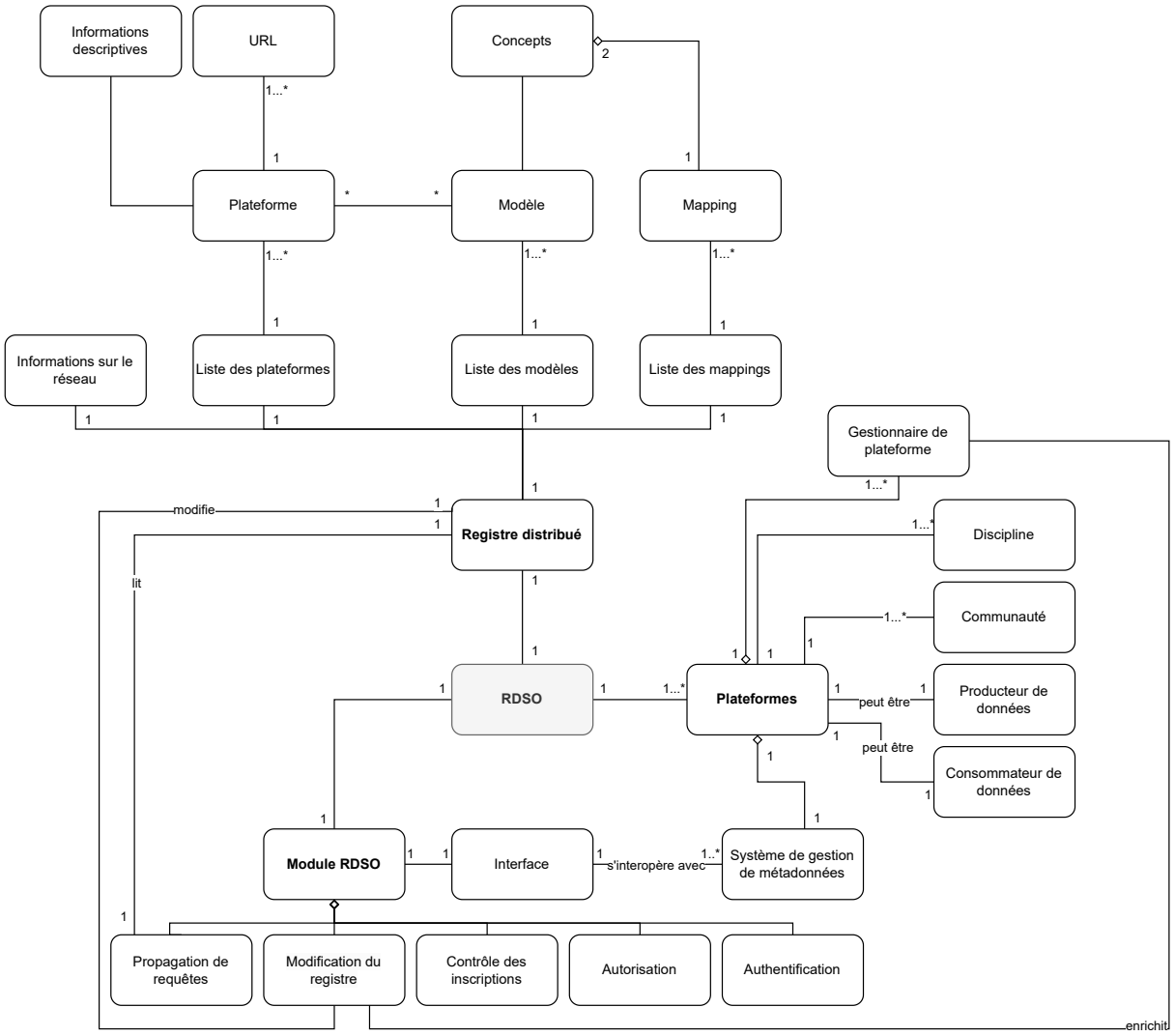

FIGURE 5.2 – Modèle de domaine du RDSO

Le fonctionnement du RDSO se base sur trois composants :
— (1) les plateformes (les PDRO)
— (2) le registre distribué
— (3) le module du RDSO
La Figure 5.2 illustre le modèle de domaine du RDSO. Ce diagramme présente les données et les composants fonctionnels (fonctionnalité, personne ou outil) du RDSO et les interactions entre ces différents objets. Pour chaque composant du RDSO que nous décrivons,

nous réalisons un agrandissement de ce modèle sur ce composant. Dans les graphes suivants, nous illustrons le composant dont nous parlons en noir et les objets externes qui interagissent avec l'élément courant en gris transparent.

5.1.1.1 Les plateformes

Le premier composant est la plateforme (cf. Figure 5.3). Dans le contexte de la Science Ouverte, ces plateformes sont les PDRO.

L'intégration d'une plateforme au RDSO est conditionnée par le fait que cette plateforme dispose des éléments suivants (cf. les relations d'agrégation dans la Figure 5.3) :

— un système de gestion de métadonnées, permettant d'exécuter une recherche dans ces métadonnées ;
— un ou plusieurs gestionnaires de plateforme, responsable de l'enrichissement du registre en utilisant la fonction de modification du registre du module RDSO.

Ces plateformes sont à la fois productrices (ouverture des données) et consommatrices (recherche de données) de données. Ces plateformes peuvent concerner une ou plusieurs disciplines et une ou plusieurs communautés.

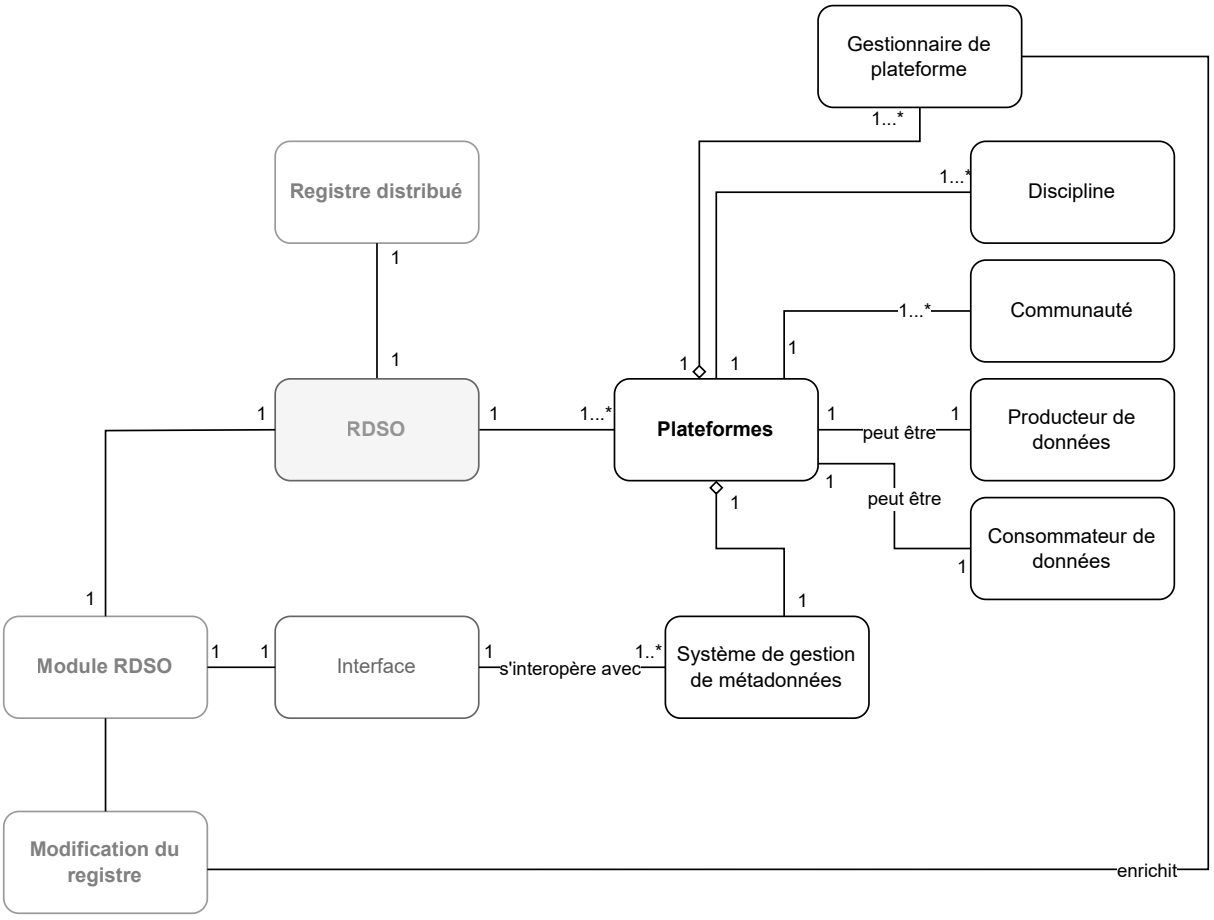

FIGURE 5.3 – Focalisation sur les plateformes dans le modèle de domaine

Les plateformes sont destinées à être autonomes dans leur gestion des données de recherche. Les outils de gestion de données ne sont donc pas considérés dans le modèle de domaine. Dans une plateforme, les modifications de ces outils n'impactent pas le fonctionnement du RDSO.

5.1.1.2 Le module du RDSO

Le module du RDSO est l'outil permettant de déployer le RDSO. Ce module est développé en Python comme un module autonome ("stand-alone"). Déployé dans une plateforme, il assure la mise en place des différentes fonctionnalités apportées par le RDSO (cf. Figure 5.4).

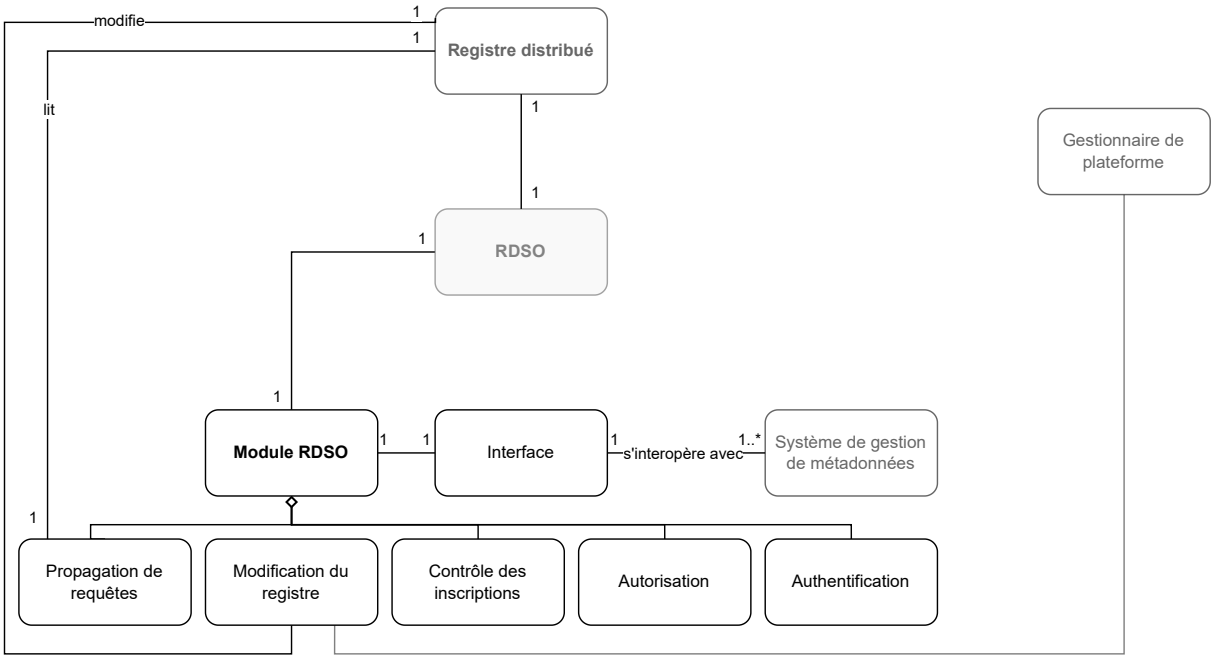

FIGURE 5.4 – Focalisation sur le module du RDSO dans le modèle de domaine

Le module du RDSO est déployé sur chaque plateforme du RDSO (cf. Figure 5.5). Il fournit une interface standardisée pour les communications entre les différentes plateformes (cf. "Interface" dans la Figure 5.4). L'interopération de cette interface passe par le développement d'une fonction Python qui accède au système de métadonnées et y exécute des requêtes. Une fois le module interopéré avec le système de gestion de métadonnées, l'ensemble des fonctionnalités du RDSO sont disponibles. Ces fonctionnalités sont les suivantes (cf. les relations d'agrégation au module RDSO sur la figure 5.4) :

— L'autorisation et l'authentification des plateformes pour les modifications du registre ;
— Le contrôle des inscriptions permettant d'apporter une résilience au réseau par un contrôle de la topologie du réseau ;
— Un mécanisme de propagation de requêtes et de modification du registre.

Nous détaillons ces différentes fonctionnalités dans la section 5.1.2.

5.1.1.3 Le registre distribué

Le RDSO se base sur un registre décentralisé répliqué entièrement sur l'ensemble des plateformes du RDSO. Toutes les plateformes conservent une copie complète du registre sous la forme d'un document JSON. L'objectif de ce registre décentralisé est d'avoir un ensemble d'informations partagées à l'ensemble des plateformes nécessaires aux fonctionnalités du RDSO.

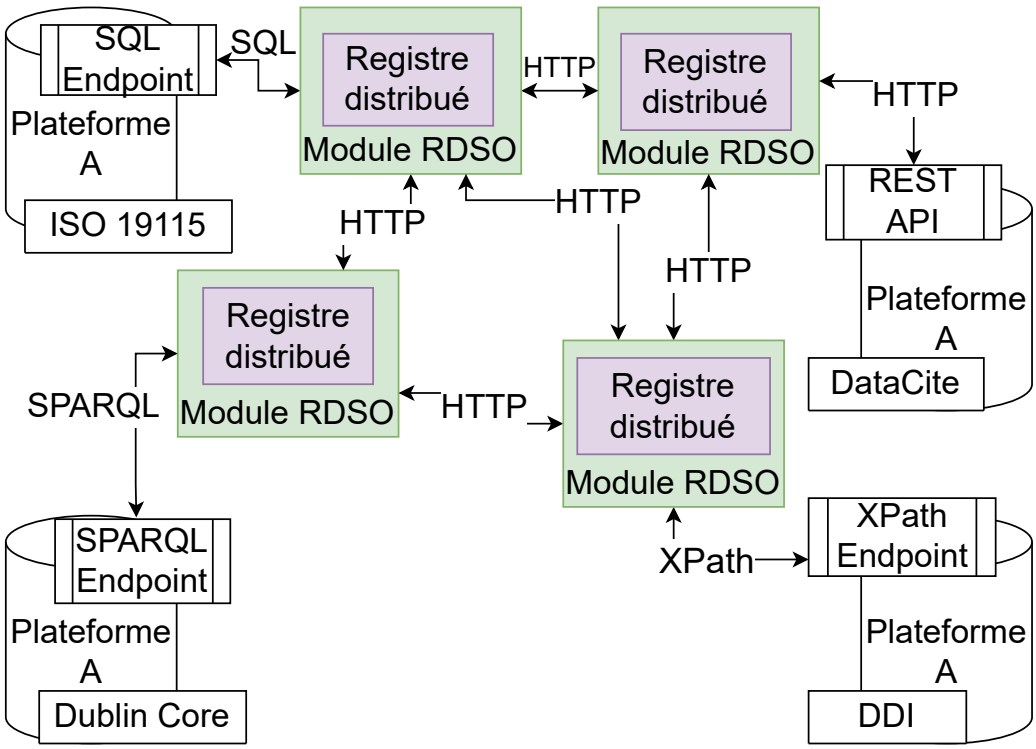

FIGURE 5.5 – Implantation du module du RDSO sur les plateformes

Ce module intègre l'ensemble des informations nécessaires au fonctionnement du RDSO (cf. Figure 5.6)

Ce registre contient quatre groupes d'informations (cf. les relations d'agrégation au registre distribué dans la figure 5.1.2.3) :

- La liste des plateformes inscrites dans le RDSO, avec des informations descriptives élémentaires de ces plateformes ;
- La liste des modèles connus dans le RDSO, contenant la liste des concepts avec un type associé ;
- La liste des mappings, chaque mapping décrivant une relation d'égalité entre deux concepts ;
- Les informations sur le réseau, avec la distribution des degrés de plateformes, afin de contrôler les inscriptions de nouvelles plateformes dans le réseau (cf. section 5.1.2.1).

Dans la suite, nous détaillons chaque groupe d'information dans le registre. Les plateformes, les modèles et les mappings sont conservés dans ce registre par un identifiant unique.

Les informations sur les plateformes permettent de renseigner les informations descriptives de la plateforme correspondante. Ces informations sont les suivantes :

- Le nom de la plateforme ;
- La liste des identifiants de voisins auxquels cette plateforme est connectée, ce qui est nécessaire pour une bonne propagation des requêtes (cf. section 5.1.2.2) ;
- La liste des URL permettant de communiquer avec le module du RDSO, nécessaire à la mise en place de requêtes (cf. section 5.1.2.2) ;
- La liste des identifiants de modèles utilisés par la plateforme ;

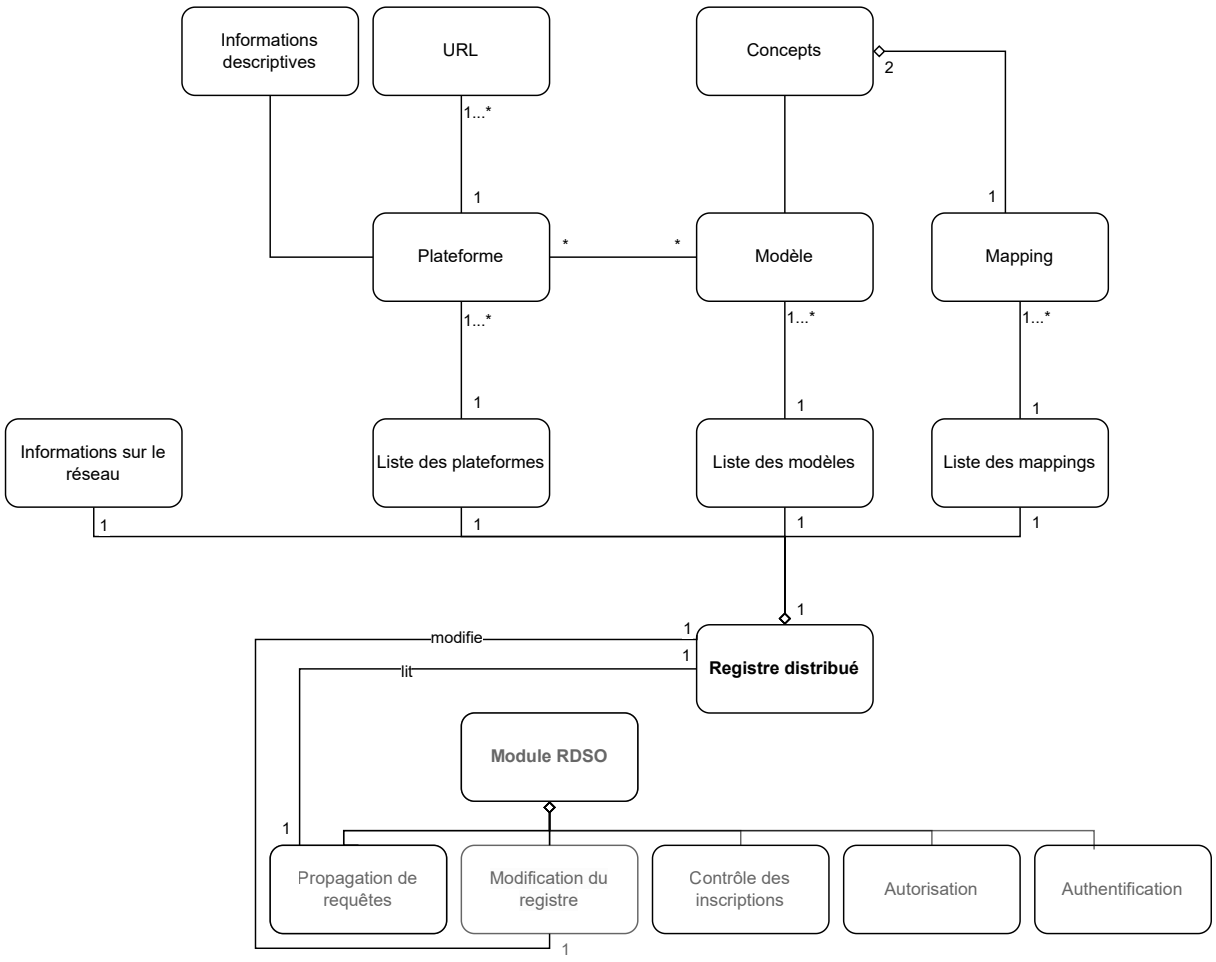

FIGURE 5.6 – Focalisation sur le registre distribué du RDSO dans le modèle de domaine

— Une clé publique de chiffrement (cf. section 5.1.2.3).

Les informations sur les modèles de métadonnées permettent de connaître les modèles utilisés dans le RDSO. Ces informations contiennent :

— Le nom du modèle ;
— La liste des identifiants des plateformes utilisant ce modèle ;
— La liste des concepts contenus dans ce modèle, prenant un modèle comme un arbre et concaténant l'ensemble des nœuds d'un chemin de la racine jusqu'à une feuille contenant un concept (cf. Figure 3.13) permettant de conserver la structure du modèle décrit. Le type de chaque concept est conservé.

La liste des types des concepts permet une reconnaissance des types des données. Cette information permet de mettre en place des mécanismes de validation des mappings en se basant sur l'idée que deux concepts contenant des types de données différents ne peuvent représenter la même information (cf. section 5.1.2.2). De plus, cette information permet d'associer aux concepts des opérations (par exemple, des opérations d'analyse sur les chaînes de caractères). Ces types sont :

— une URL ("URL") ;
— un URI ("URI") ;
— un identifiant ("id") ;
— une chaîne de caractères ("string") ;

— une date ("date") ;
— un mail ("mail") ;
— un nombre ("number") ;
— une liste pouvant contenir des types ("list[type]", avec une syntaxe similaire au langage Java [1]).

Les informations sur les mappings entre les concepts contenus dans le registre contribuent à l'interopérabilité en mettant en place des passerelles entre modèles. Chaque mapping est indexé par un identifiant unique et permet de définir une relation d'égalité entre deux concepts. Chaque mapping est défini par les éléments suivants :

— Les clés des deux concepts associés (mappés) provenant de la liste des concepts des modèles ;
— Les identifiants des modèles associés ;
— La liste des plateformes ayant validé le mapping ;
— Le score de vraisemblance du mapping ;
— Le type de mapping, qui peut être manuel ou automatique.

Ces mappings sont nécessaires au mécanisme d'échange d'informations.

Les informations sur le réseau contiennent la distribution des degrés des plateformes dans le réseau. Cette distribution prend la forme d'une liste de listes. Chaque liste imbriquée conserve les identifiants des plateformes ayant un degré égal à son index dans la liste, avec un degré minimal de deux (cf. Section 5.1.2.1). Ces informations permettent d'obtenir une vision globale du réseau et sont nécessaires pour mettre en place le mécanisme d'inscription (cf. Section 5.1.2.1).

Ce registre est enrichi par les gestionnaires de chaque plateforme, au moment de l'inscription (cf. section 5.1.2.1) avec les informations descriptives de leur plateforme, le(s) modèle(s) utilisés par leur plateforme ainsi que des mappings entre leurs modèles et d'autres présents dans le registre. Les gestionnaires modifient aussi le registre tout au long du fonctionnement du RDSO pour l'ajout de mappings, une mise à jour des informations ou une revue des mappings existants (cf. section 5.1.2).

Gestion des écritures concurrentes Ce registre est répliqué dans son intégralité à l'ensemble des plateformes. Le problème écritures concurrentes est résolu par trois actions :

— En restreignant l'ajout, la modification et la suppression des informations de plateformes, aux plateformes ayant renseigné ces informations ;
— En empêchant la suppression et la modification de modèles de métadonnées, en considérant qu'une modification génère une nouvelle version du modèle de métadonnées utilisable par d'autres plateformes ;
— En conservant l'ensemble des mappings.

Pour éviter d'avoir des mappings pouvant être contradictoires et valider la fiabilité des mappings, chaque mapping comporte un score de vraisemblance est défini comme suit :

$$Score = Q(T + Rev)$$

avec

— La présence d'un type équivalent des deux concepts, Q ;
— Le type de match (manuel ou automatique), T ;
— La somme des avis donnés (+1 ou -1) par d'autres utilisateurs, Rev.

1. `https://docs.oracle.com/javase/8/docs/api/java/util/List.html`

La présence d'une réponse à une requête utilisant ce mapping permet d'avoir un premier niveau de nettoyage des mappings afin d'ignorer ceux mis en place sur deux entités qui ne possèdent pas le même type de contenu, par extension la même information. Par exemple, une date et une position géographique ne peuvent être utilisées par les mêmes opérations. Nous considérons qu'un mapping manuel est réalisé grâce aux connaissances expertes, et donc nous considérons les mappings manuels ($T = 1$) plus fiables que les mappings automatiques ($T = 0$). Ensuite, les avis donnés peuvent indiquer si le mapping est correct (d'une valeur de +1) ou non (d'une valeur de -1). La somme de ces scores permet de comparer les mappings. Une sélection de N mappings, triés par ordre de vraisemblance, est choisie parmi les mappings portant sur le même concept.

Les mappings utilisés sont les mappings les plus vraisemblables (c'est-à-dire ceux avec le score de vraisemblance le plus élevé). Lorsqu'il lance une recherche, l'utilisateur fixe un seuil pour savoir combien de mappings doivent être conservés pour cette requête. La présentation des résultats de requêtes peut être réalisée par ordre de vraisemblance.

5.1.2 Fonctionnalités du RDSO

Les fonctionnalités du RDSO sont définies pour permettre l'utilisation des composants définis dans la section précédente et réaliser l'objectif de recherche unifiée de données. Plusieurs critères doivent être réunis pour avoir une solution pérenne et fonctionnelle :

— Mettre en place un mécanisme d'échange d'information et de données ;
— Avoir un réseau résilient face aux pannes et aux attaques ciblées pour éviter une scission du réseau en deux, qui empêcherait l'échange entre les différentes composantes connexes ;
— Mettre en place des mécanismes de sécurité pour éviter les falsifications d'informations ou les usurpations d'identité, qui empêcheraient le réseau de fonctionner.

Pour la résilience et le passage à l'échelle du réseau, nous avons choisi une topologie de réseau d'invariant d'échelle (Barabási and Pósfai (2016)), qui assure une résilience face aux pannes et aux attaques. Cette topologie possède des caractéristiques pour le passage à l'échelle de l'échange d'information et de données (Barabási and Pósfai (2016)). Pour s'assurer de la mise en place de cette topologie, nous intégrons un mécanisme de contrôle des inscriptions dans le réseau (cf. section 5.1.2.1).

Pour assurer l'échange d'information et gérer la mise à jour du registre distribué à l'ensemble des plateformes du RDSO, nous avons intégré au RDSO un mécanisme de propagation des requêtes et des modifications de registre, permettant une recherche de données de recherche sur l'ensemble des plateformes du RDSO (cf. section 5.1.2.2)

Enfin, afin de garantir la sécurité des recherches et pour éviter toute modification malicieuse du registre, les interactions dans le RDSO se basent sur un mécanisme d'authentification basé sur le chiffrement (cf. section 5.1.2.3).

5.1.2.1 Résilience du réseau

Pour s'assurer d'une résilience aux pannes, aux attaques pouvant scinder le réseau en deux composantes non connexes qui empêcheraient son bon fonctionnement, la topologie du réseau doit être prise en compte (cf. la théorie de la percolation dans la science des réseaux (Barabási and Pósfai (2016)).

Dans les données de Re3Data, nous avons observé la distribution des degrés dans les plateformes de la Science Ouverte. Une grande quantité de plateformes possède un degré faible et le nombre de plateformes décroit de manière exponentielle en fonction du

degré de ces plateformes (cf. Figure 3.12). Ce même phénomène est présent dans d'autres réseaux comme Internet, le World Wide Web (WWW) ou les collaborations scientifiques (Barabási and Pósfai (2016)).

Ce type de réseau est un réseau invariant d'échelle (ou réseau sans échelle ou "scale-free network"). Ce type de réseau possède une spécificité : la présence de hubs. Les hubs sont des nœuds du graphe possédant un grand nombre de connexions, comme par exemple les serveurs de Google dans le WWW ou des articles hautement cités dans les collaborations scientifiques. De plus, dans ce type de réseau, la distribution des degrés suit une loi de puissance de paramètre λ. Par souci de simplification, nous utiliserons dans la suite l'appellation *réseaux invariants d'échelle de paramètre* λ pour parler les réseaux invariants d'échelle suivant une loi de puissance de paramètre λ. Ce type de réseau possède plusieurs propriétés (Barabási and Pósfai (2016)) :

- "Ultra small world" - Très petit monde : Les réseaux invariants d'échelle de paramètre $2 < \gamma < 3$ possèdent une croissance du diamètre de réseau en $ln(ln(n))$, où n est le nombre de nœuds du réseau. Cela indique que les hubs permettent de grandement réduire la distance à parcourir dans le réseau, notamment dans le cas d'un échange d'information.
- Une approche épidémiologique appliquée au réseau montre que le temps de propagation d'une information avec un réseau invariant d'échelle de paramètre $\gamma \leq 3$ tend vers 0. Autrement dit, la propagation d'une information tend à être instantanée.
- La robustesse face aux attaques et aux pannes simultanées augmente grandement dans un réseau invariant d'échelle à partir d'un degré moyen de 3. Plus le degré moyen augmente, plus la robustesse est grande.

Cette topologie vise à organiser les échanges entre les plateformes du réseau, avec le mécanisme de propagation de requêtes mais aussi le mécanisme de propagation des informations du registre distribué. Cette topologie permet ainsi d'avoir des propriétés adaptées à la Science Ouverte pour répondre à cette problématique d'échange d'information. Les propriétés de passage à l'échelle, de vitesse de propagation d'une information au sein du réseau et les aspects de robustesse de cette topologie sont adaptés au contexte de la Science Ouverte. De plus, cette topologie est déjà en place dans le réseau de collaboration scientifique et semble être suivie par les plateformes de gestion de données (cf. Figure 3.12), montrant une adéquation de cette topologie à la Science Ouverte. Nous choisissons d'utiliser cette topologie pour le RDSO pour ces différentes raisons.

Algorithme de contrôle des inscriptions Pour permettre au RDSO de se déployer en suivant cette topologie, nous avons ajouté un mécanisme de contrôle des inscriptions. Ce mécanisme est décrit par le pseudo-algorithme 1. Nous avons choisi de mettre en place un réseau invariant d'échelle de paramètre $\gamma = 2.5$. Donc nous avons bien $2 > \gamma > 3$ et nous évitons les potentiels effets de bords en choisissant la valeur centrale de cet intervalle. Nous avons choisi la valeur 2 pour le nombre minimal obligatoire de voisins à chaque plateforme afin de réduire le coût d'inscription. Mais nous recommandons fortement de connecter une nouvelle plateforme à 3 voisins minimum.

Cet algorithme permet de trouver le nœud auquel est connectée la nouvelle plateforme, telle que la nouvelle distribution générée soit la plus proche possible de la distribution théorique que nous souhaitons, à savoir une topologie de réseau invariant d'échelle de paramètre $\gamma = 2.5$.

La fonction "add_edge" ajout un arc entre deux nœuds dans le graphe du réseau.

Algorithm 1: Inscription d'une nouvelle plateforme dans le RDSO

Input : Un graphe du réseau du LDSO $G = (V, E)$, un nœud représentant la plateforme s'inscrivant $v_{inscrit}$, un nœud sélectionné par l'utilisateur v_{choisi}

Output: Un graphe réseau $G' = (V', E')$

```
if |V| = 0 then
    add_node_to_network(G, v_inscrit)
if |V| = 1 then
    add_node_to_network(G, v_inscrit)
    add_edge(v_inscrit, v_0) /* v_0 le seul nœud du réseau */
v if |V| ≥ 2 then
    add_node_to_network(G, v_inscrit)
    add_edge(v_inscrit, v_choisi)
    v_plus_proche = get_node_nearest_from_distribution(v_inscrit, V) /* Retourne
       le nœud auquel se connecter qui permet à la distribution de degré de se
       rapprocher le plus d'une distribution suivant une loi de puissance de
       paramètre γ = 2.5 */
    add_edge(v_inscrit, v_plus_proche)
return G
```

La fonction "add_node_to_network" ajoute un nœud dans le graphe du réseau.

La fonction *get_node_nearest_from_distribution* retourne un nœud de degré k_{opti} tel que k_{opti} minimise la divergence de Kullback-Leibler D_{KL}, donnant une mesure de dissimilarité entre la distribution de degré actuelle dans le RDSO et une distribution suivant une loi de puissance de paramètre $\gamma = 2.5$. Nous sélectionnons le premier nœud de la liste des nœuds de degré k_{opti} pour réduire le temps de calcul dans notre implantation.

Pour comprendre en profondeur, nous définissons P, la distribution des degrés dans le graphe du réseau initial avec le nœud v_{choisi} ajouté dans le graphe sans être connecté, P_n la distribution des degrés dans le graphe du réseau après la connexion de v_{choisi} à un nœud de degré n, Q une loi de puissance de paramètre $\lambda = 2.5$, E_k l'ensemble des nœuds de degrés k dans la distribution P et E l'ensemble des nœuds de la distribution P. Nous supposons que v_{choisi} possède un degré de 2, avec des arêtes connectant v_{choisi} à lui-même (des boucles). L'objectif de cet algorithme est de remplacer ces boucles par des connexions à des plateformes. Nous avons :

$$D_{KL}(P||Q) = \sum_{k \in K} q(k) \log(\frac{p(k)}{q(k)})$$

avec k_{max} le degré maximal de la distribution P, K l'ensemble des degrés possibles dans ce graphe, de 2 à k_{max}, $p(k)$ et $q(k)$, respectivement, la fonction de masse de la distribution P et Q. La connexion de v_{choisi} à un nœud de degré k augmente le degré du nœud auquel on connecte le nœud v_{choisi} de 1. Cet ajout enlève un nœud de degré n et ajoute un nœud de degré $n+1$ à la distribution. Ainsi, nous avons

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

avec $p_n(k)$ la probabilité empirique d'avoir un nœud de degré k dans la distribution P_n. $p_n(k)$ est indépendant de n quand $k \notin \{n, n+1\}$. Ceci pose un problème d'optimisation décrit avec l'équation (5.1). Nous avons $q(n) = \frac{n^{-\gamma}}{\zeta(\gamma)} = \frac{n^{-2.5}}{\zeta(2.5)}$, avec ζ la fonction Zeta de Riemann Barabási and Pósfai (2016). Comme la valeur de $\zeta(2.5)$ est indépendante de n, il est possible de simplifier l'équation 5.1 en équation 5.2.

$$k_{opti} = \arg\min_n(p_n(n)\log(\frac{p_n(n)}{q(n)}) + p_n(n+1)\log(\frac{p_n(n+1)}{q(n+1)})) \tag{5.1}$$

$$k_{opti} = \arg\min_n(\frac{|E_n|-1}{|E|}\log(\frac{\frac{|E_n|-1}{|E|}}{n^{-2.5}}) + \frac{|E_{n+1}|+1}{|E|}\log(\frac{\frac{|E_{n+1}|+1}{|E|}}{(n+1)^{-2.5}})) \tag{5.2}$$

L'équation 5.2 donne le degré optimal du nœud à connecter dans la fonction *get_node_nearest_from_dis* Nous évaluons n allant de 2 à k_{max}. La complexité de calcul de cet algorithme a une complexité en $O(k_{max})$.

Nous observons que le RDSO est un réseau d'overlay maillé, pouvant se définir intuitivement comme un réseau fonctionnant sur un autre réseau (cf. la définition précise de réseau d'overlay proposée par Van Steen and Tanenbaum (2023)) - dans notre cas le réseau Internet - proposant plusieurs chemins pour un même nœud (cf. la définition précise par Van Steen and Tanenbaum (2023)), offrant une robustesse par la présence de plusieurs chemins différents pour accéder au même nœud.

Combiné à la topologie choisie, le RDSO propose une robustesse accrue face à des délétions volontaires (c'est-à-dire les cyberattaques) ou involontaire (c'est-à-dire les pannes) de nœuds du réseau et un temps de propagation des informations réduit. La section suivante décrit en détail cette propagation des informations, à savoir des requêtes mais aussi des modifications du registre.

5.1.2.2 Mécanisme de propagation

Ce mécanisme de propagation permet, grâce aux réponses à l'interopérabilité des plateformes de gestion de données de la Science Ouverte, de réaliser une recherche transparente de données de recherche. De plus, ce même mécanisme est utilisé pour les mises à jour du registre distribué.

Propagation des requêtes Lors d'une requête sur une plateforme (cf. Figure 5.7), cette requête est envoyée aux voisins de la plateforme courante ("plateforme locale"). Une fois la requête reçue sur une plateforme voisine (plateforme P1 ou P2), la liste des mappings est récupérée, afin de transformer la requête exprimée avec le modèle de la plateforme locale en une requête exprimée avec le modèle utilisé sur la plateforme P2. Une fois exécutée sur P2, la requête est ensuite envoyée aux voisins de la plateforme P2 et traduite à nouveau à l'aide des mappings. Ce processus est répété jusqu'à ce que la requête soit reçue par l'ensemble des plateformes du réseau (P3 et P4), par inondation du réseau. Les résultats sont ensuite agrégés tout au long du chemin dans une réponse unique fournissant l'ensemble des réponses trouvées sur toutes les plateformes interrogées.

La gestion des requêtes par une plateforme, qu'elle soit locale (c'est-à-dire réalisée par un utilisateur de cette plateforme) ou provenant d'une autre plateforme, est définie par l'algorithme 2.

La fonction "remplace_concept" remplace la chaîne de caractères d'un concept par la chaîne de caractères d'un autre concept.

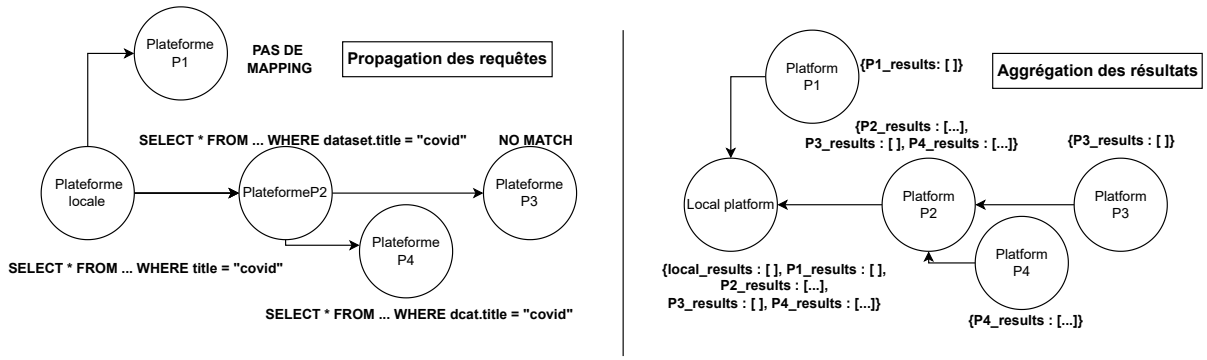

FIGURE 5.7 – Illustration du mécanisme de propagation de requête

La fonction "ajout" ajoute un identifiant de requête à une liste d'identifiants.

La fonction "requete_voisin" envoie la requête aux voisins de la plateforme courrante.

La fonction "get_metadata" permet de récupérer les métadonnées auprès du système de métadonnées local en utilisant une requête.

Lors de la réception d'une requête (locale ou provenant d'un voisin) et pour éviter d'exécuter deux fois la même requête, chaque identifiant de requête est inséré dans une liste servant à ignorer une requête déjà traitée (cf. lignes 1 à 3 dans l'algorithme 2).

Si la requête est nouvelle, l'identifiant de la requête est ajouté à cette liste puis la liste des concepts de la requête est remplacée par la liste des concepts correspondants dans le modèle local (cf. ligne 4 à 6 dans l'algorithme 2).

Cette requête est exécutée dès qu'un seul concept est mis en correspondance avec le modèle cible, ce qui peut résulter dans une sous-requête ne s'appliquant pas sur la totalité des concepts initiaux (cf. lignes 7 à 8 dans l'algorithme 2).

Une fois la requête executée, la plateforme attend les résultats de la requête fournis par ses voisins, puis retourne un document contenant les résultats des autres plateformes et ceux obtenus sur elle-même (cf. lignes 9 à 10 dans l'algorithme 2).

Le nombre de concepts initiaux mappés et interrogés représente un niveau d'interopérabilité des modèles pour une requête spécifique (par exemple, si un seul des deux concepts composant une requête est mappé, alors les modèles ont un niveau d'interopérabilité de 50% pour cette requête). Une présentation des informations peut être faite en fonction du niveau d'interopérabilité entre les modèles, puis en fonction de la vraisemblance des mappings utilisés, selon les besoins des utilisateurs. Cette approche permet d'assurer la plus grande richesse de données possible pour les utilisateurs.

Ce mécanisme permet de trouver les jeux de données là où ils se trouvent, sans modification des outils de gestion de données disponibles sur une plateforme grâce à l'utilisation des mappings et du registre distribué.

Modification du registre Les modifications sur le registre doivent aussi être propagées quand elles sont effectuées. L'algorithme utilisé est le même que pour la propagation des requêtes. L'exécution des requêtes est remplacée par un ajout des modifications effectuées. Aucune résolution de conflit n'est nécessaire, grâce aux contraintes sur les opérations sur le registre.

Comme pour les requêtes, l'approche consiste à inonder le réseau avec les nouvelles informations. Les écritures concurrentes ont été évitées par des restrictions sur les opérations de modifications du registre. Ainsi, les modifications validées sont envoyées aux voisins et

Algorithm 2: Mécanisme de propagation de requête

```
Input  : Une requête Q
Output: Une liste de résultats R
if Q.identifiant in liste_requete_traitees then
    ignorer() ;
ajout(Q.identifiant, liste_requete_traitees) ;
for concept in Q do
    if ((concept_Q, concept_externe) in registre AND concept_externe in
     modele_local) then
        local_Q = remplace_concept(Q, concept, concept_externe) ;
        mapping_trouve = Vrai ;
if mapping_trouve then
    resultats_locaux = get_metadata(local_Q) ;
resultats_externes = requete_voisin(Q, registre) ; /* Envoi de la requête aux
    voisins, grâce aux informations contenues dans le registre, et attend le
    retour de cette requête                                                  */
R = aggregation(resultats_locaux, resultats externes) ; /* Agrège les résultats en
    un seul document JSON, où chaque clé détermine les résultats d'une
    plateforme                                                               */
return R
```

la réception de plusieurs modifications en simultané résulte en une fusion des différentes modifications.

5.1.2.3 Authentification par chiffrement

L'authentification et l'autorisation des écritures, notamment sur les modifications d'informations de plateforme, sont réalisées via l'utilisation de clés de chiffrement. Une clé publique de chiffrement est fournie par chaque plateforme à l'inscription. Cette paire de clés assure la signature numérique des messages envoyés par une plateforme. Soit la plateforme A, souhaitant s'inscrire. Cette plateforme génère un couple de clés publique / privée, C_{pub} et C_{priv}, à partir de l'algorithme AES 256 (Daemen and Rijmen (1999)). Cet algorithme de chiffrement est recommandé par la NSA [2] notamment pour sa sûreté. La plateforme A inscrit dans ses informations sa clé publique et conserve sa clé privée (cf. Figure 5.8).

Lors d'une modification d'une partie restreinte par authentification, la plateforme A chiffre le message avec sa clé privée C_{priv}. La plateforme B, recevant le message, déchiffre le message en utilisant la clé publique C_{pub} contenue dans le registre, permettant de s'assurer de l'identité de l'émetteur du message (cf. Figure 5.9). Toute autre clé que la clé C_{priv} pour chiffrer le message résultera en une erreur de déchiffrement, ce qui correspondrait à une tentative d'usurpation d'identité. La sécurité est déléguée aux plateformes, qui ont pour objectif de s'assurer de la confidentialité de leur clé privée C_{priv}.

Seules les modifications du registre de plateformes sont soumises à autorisation. Cependant, le mécanisme d'authentification peut être utilisé pour sécuriser l'ensemble des

2. https://www.ibm.com/docs/fr/ibm-mq/9.3?topic=tls-national-security-agency-nsa-suite-b-cryptography

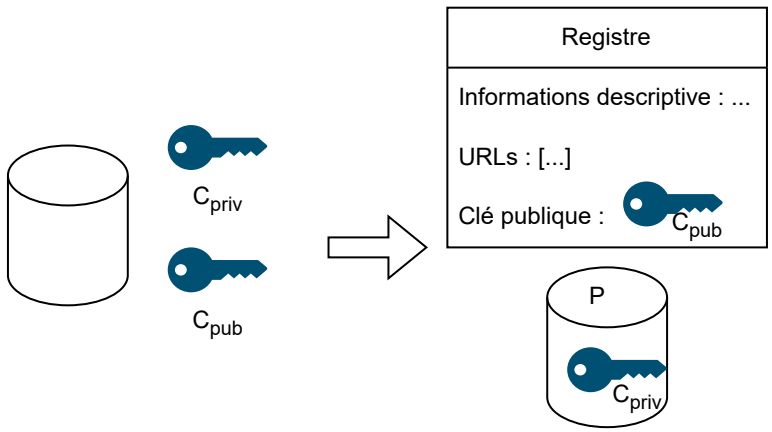

FIGURE 5.8 – Création de la paire de clés de chiffrement par la plateforme P

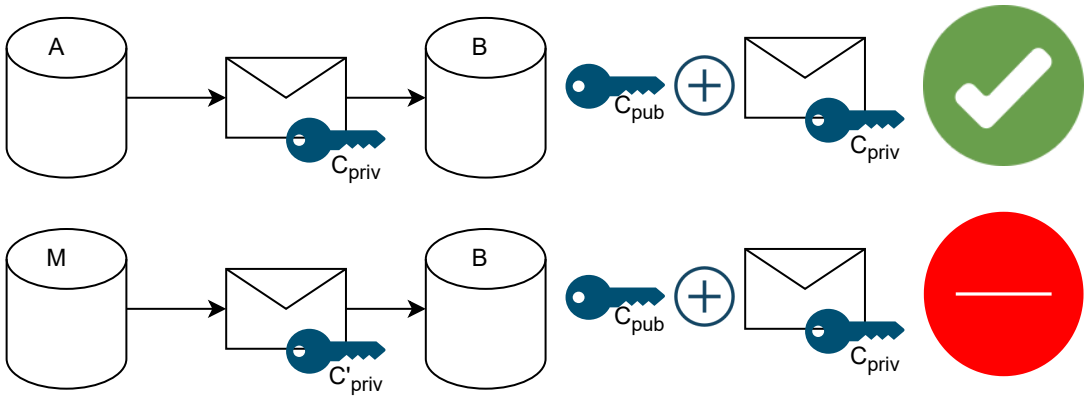

FIGURE 5.9 – Signature d'un message par la plateforme A avec sa clé privée C_{priv} et tentative d'usurpation d'identité par la plateforme M avec une clé C'_{priv}

communications du réseau, notamment pour éviter des attaques par déni de service (DoS) ou pour le partage de ressources de calcul des plateformes entre certaines entités d'une même communauté.

5.2 Implantation de l'interopérabilité des PDRO par le RDSO

Pour la mise en place de l'interopérabilité, les deux hétérogénéités à gérer sont (1) l'hétérogénéité des API de communication et (2) l'hétérogénéité des modèles de métadonnées. Pour proposer une recherche interdisciplinaire et intercommunautaire de données à l'ensemble des PDRO, le problème du passage à l'échelle des mécanismes est un des verrous à lever.

5.2.1 La variété d'API de communication

Pour répondre à l'hétérogénéité des API de communication, nous avons opté pour un mécanisme d'interopérabilité par standardisation. Nous avons défini une API REST standard pour l'ensemble des plateformes du RDSO qui définit les mécanismes de communications entre ces plateformes. Cette standardisation est possible car il n'existe à l'heure

actuelle aucune solution à l'échelle de la Science Ouverte similaire. Aucune API déjà en place ne permet cet échange ce qui nécessite une interopération avec le RDSO.

De plus, l'utilisation d'API REST permet une flexibilité de développement et facilite la mise en place de mécanismes d'interopérabilité grâce à des passerelles en plus de la standardisation.

5.2.2 La variété des modèles de métadonnées

Pour gérer l'hétérogénéité des modèles de métadonnées, nous avons choisi de mettre en place des passerelles entre les modèles, qui sont les mappings. Chaque plateforme utilise le ou les modèles qui lui sont nécessaires. Ces modèles sont décrits dans le registre et les concepts de ces modèles sont alignés entre eux pour fournir un accès transparent à l'ensemble des métadonnées des plateformes du RDSO.

5.3 RDSO et maille de données

Les mailles de données (traduit de l'anglais "data mesh") illustrent un changement de paradigme dans la conception d'architecture de gestion de données orientées Big Data dans les entreprises. Les mailles de données visent une approche distribuée de la gestion de données, en opposition avec l'approche centralisée des lacs de données Dehghani (2019). L'architecture logique des mailles de données se base sur quatre principes Dehghani (2020) :

— La propriété du domaine ("*Domain Ownership*") : La responsabilité de la gestion des données est décentralisée et distribuée aux personnes les plus proches des données, à savoir les personnes proches des processus de création de données dans les différents domaines de l'entreprise (ressources humaines, comptabilité, finance, analyse de données, etc...).

— Les données comme un produit ("*Data as a product*") : Les données doivent être considérées comme des produits et les consommateurs sont des clients de l'entreprise. La gestion des données intègre la qualité de données, la réponse aux silos de données et permet l'utilisation autonome des données par les clients. Un quantum architectural est défini avec les données-produits ("Data product"), définissant un nœud dans la maille de données permettant la gestion de données et des métadonnées associées, des pipelines de traitements de données ainsi que l'infrastructure garantissant cette autonomie.

— Les plateformes en libre-service ("*Self-serve data platform"*) : Les infrastructures existantes doivent offrir un plus grand panel de fonctionnalités et d'outils pour gérer et créer de nouveaux pipelines de traitement de données à un coût réduit, notamment à travers l'utilisation de stockage polyglotte de données extensible, des schémas de données-produits ou d'outils d'orchestration de pipelines de traitements.

— Une gouvernance informatique fédérée ("*Federated computational governance*") : La gouvernance des données est un équlibre entre une gouvernance locale (par les données-produits) et une gouvernance globale (par la plateforme de la maille de données) pour répondre à la question de l'interopérabilité des différentes données-produits. Cette gouvernance est la décentralisation et l'autosouveraineté des domaines, l'interopérabilité basée sur une standardisation à l'ensemble de la maille

de données, une topologie dynamique et l'exécution automatisée des décisions par la plateforme.

Pour permettre une comparaison, nous considérons les deux architectures comme des réseaux. Ces deux réseaux visent à rendre la responsabilité et le contrôle des données au plus près des créateurs de données pour fluidifier la gestion des données au sein du réseau. Nous avons comparé le RDSO et la maille de données (cf. tableau 5.3) et nous observons les similitudes suivantes entre les deux :

— la présence d'une décentralisation ;
— la possibilité d'interopération entre les entités de l'architecture ;
— une autonomie des plateformes.

Cependant, ces deux réseaux diffèrent en plusieurs points :

— **Objectifs différents :** Le RDSO en l'état ne vise qu'un échange d'information. La maille de données a pour objectif de réaliser des analyses de données avec des sources hétérogènes locales et extérieures. La maille de données vise donc une étape supplémentaire par rapport au RDSO. Cependant, un enrichissement de l'objectif est possible.
— **Quantum architectural :** Le quantum architectural de la maille de données est la "donnée-produit". Le réseau est conçu autour de la donnée. Le quantum architectural du RDSO est la plateforme. Même si l'objectif final est de permettre aux plateformes d'échanger des informations et des données de recherche, le RDSO se concentre sur les plateformes et propose des solutions exploitant ces plateformes et non des données.
— **Portée de l'interopération :** La maille de données est construite à partir d'une philosophie mettant le domaine au centre de la réflexion. Les domaines possèdent des "données-produits" et l'objectif recherché est une interopération de ces domaines (similaires aux disciplines) et de leur "données-produits" mais au sein d'une même organisation ou entreprise (communauté). Donc l'interopération est interdisciplinaire et intracommunautaire. Le RDSO propose une interopération interdisciplinaire, intercommunautaire mais peut répondre aussi à une intradisciplinarité et intracommunautarité.
— **Gouvernance :** Enfin, la maille de données met en place une gouvernance vue comme un groupement fédéré des différents domaines pour assurer l'interopérabilité par standardisation (modélisations, qualité des données, sécurité de données, surveillances). Cette gouvernance fédérée correspond à une prise de décision centralisée (Dehghani (2019)). Cependant, des gouvernances fédérées non centralisées ont été pensées, pour pallier les problèmes liés à la centralisation. De plus, l'interopération des différentes plateformes se base dans un contexte d'une architecture technique commune partagée au sein d'une même entreprise. Le RDSO se place dans un contexte ne permettant pas la centralisation et la standardisation des modélisations. Il n'y a pas de décision globale et aucune autorité ne se place au-dessus des plateformes. Ainsi, toutes les plateformes sont totalement autonomes et le RDSO ne fixe aucune gouvernance globale.

Le concept de maille de données est un concept récent avec de nombreuses évolutions. Wider et al. (2023) proposent une approche décentralisée de cette gouvernance. Cependant, cette décentralisation n'est pas complète et reste sur le même paradigme initial d'une entité qui prend les décisions de gouvernance globale prioritaires sur les décisions locales, notamment grâce à des mécanismes de vérification de l'autorité centrale.

Dolhopolov et al. (2023c) enrichissent l'architecture de maille de données avec l'ajout

	RDSO	Maille de données
Contexte	Science Ouverte	Entreprise
Architecture	Décentralisée	Décentralisée
Contrôle des données	Locale	Locale
Gouvernance	Pas de gouvernance globale, chaque plateforme s'autogouverne	Hybride (centralisée & décentralisée)
Objectif	Echange de données	Consommation de sources de données hétérogènes
Quantum architectural	Plateforme	Données-produit
Philosophie	Plateforme au centre	Domaine au centre
Portée de l'interopération	Intradisciplinaire, intracommunautaire, interdisciplinaire et intercommunautaires	Interdisciplinaire et intracommunautaire

d'une gestion distribuée des catalogues de métadonnées et des catalogues des produits d'une plateforme de gestion de données. Les auteurs présentent une implémentation de ces catalogues distribués à travers l'utilisation des technologies de blockchains (Dolhopolov et al. (2023b) et Dolhopolov et al. (2023a)). Dans le RDSO, la gestion des catalogues de métadonnées est décentralisée. Le registre distribué permet de conserver les informations nécessaires au requêtage des différents catalogues mais les catalogues de métadonnées ne sont pas partagés entre les plateformes.

Cependant, les auteurs affirment que la gouvernance doit être non centralisée, allant à l'encontre de la définition initiale de la maille de données (Dehghani (2019)). Cette décentralisation permet notamment une application à la Science Ouverte, bien que l'approche de l'interopérabilité par standardisation ne soit pas compatible avec la Science Ouverte.

Même si le RDSO possède des similitudes avec une maille de données, leurs approches de l'interopérabilité sont totalement opposées et incompatibles. Le même constat est réalisé sur la gouvernance.

Le RDSO est un réseau d'overlay maillé, ce qui explique les similitudes avec les mailles de données, mais n'est pas une maille de données, si on se réfère aux travaux actuels sur les mailles de données. Cette architecture est jeune et évolue beaucoup dans la littérature, pouvant, à l'avenir, permettre de rentrer le RDSO dans les critères de définition d'une maille de données.

5.4 Conclusion

La recherche de données interdisciplinaire permet un enrichissement des projets de recherche, en enrichissant des approches et des outils d'analyses proposés aux chercheurs. Les contraintes apportées, en termes de volumétrie (de plateformes, des données) et de variété (de modélisation de métadonnées), ne permettent pas à des solutions centralisées de voir le jour pour répondre à la problématique. Nous avons proposé une solution de réseau d'interconnexion de plateformes décentralisé, distribué et fédéré avec le Réseau de Données de la Science Ouverte. Cette solution est basée sur un module autonome à intégrer au sein de chaque plateforme du réseau, sans avoir à modifier la plateforme, ainsi qu'un registre distribué et répliqué dans son ensemble à la totalité des plateformes du RDSO, contenant les informations d'interopération des modèles de métadonnées. L'intégration de ce module passe par la modification d'une fonction de connexion du module au système de gestion des métadonnées. Cette solution se base sur une propagation des requêtes et des informations du registre, pour permettre de trouver les informations où elles se trouvent sans nécessiter le stockage local des données.

Le RDSO se base sur des hubs communautaires. Ces nœuds du graphe possèdent un grand nombre de connexions. Un volume plus élevé que les autres plateformes est attendu sur ces hubs. L'intégration du LDSO au sein du RDSO, comme une solution répondant aux besoins de passage à l'échelle des hubs communautaires et intégrant des outils adaptés à la Science Ouverte, est pensée pour permettre une synergie des deux solutions proposées.

Pour implanter l'interopérabilité des PDRO, le RDSO se base sur :

— La standardisation des API de communication qui peut être enrichie si besoin par des mécanismes d'interopérabilité grâce à des passerelles basées sur des API REST ;

— Des passerelles pour les modèles de métadonnées, grâce à l'utilisation de mappings.

En l'état, les mappings servent à définir des relations d'égalité entre les concepts. La possibilité de définir d'autres relations sémantiques entre les concepts permettrait un enrichissement du mécanisme de recherche de données. Nous discutons de cette perspective d'évolution dans le chapitre 7.

Nous évaluons le RDSO dans le chapitre 6, afin de valider les capacités d'échange d'information mais aussi pour quantifier les différents apports du RDSO dans la Science Ouverte.

Le RDSO a fait l'objet de deux publications scientifiques dans la conférence internationale RCIS 2024 (Dang et al. (2024a)) et dans la conférence internationale DASFAA 2024 (Dang et al. (2024b)).

Chapitre 6

Expérimentations et validations

La Science Ouverte est définie comme “un accès ouvert aux connaissances partagées et développées à travers un réseau de collaborations dans le monde de la recherche scientifique” (Vicente-Saez and Martinez-Fuentes (2018)). Les collaborations passent par une recherche de données intercommunautaire, intracommunautaire et interdisciplinaire. Les données sont disséminées sur une grande variété de PDRO. Il est donc nécessaire de faciliter l'échange de métadonnées pour permettre une recherche unifiée sur ces PDRO, qui est demandée par les chercheurs.

Le partage de métadonnées à l'échelle de la Science Ouverte est géré à deux niveaux :

— Au niveau local, avec l'échange de métadonnées intracommunautaire ;
— Au niveau global, avec l'échange de métadonnées intercommunautaire et interdisciplinaire.

Nous avons proposé deux solutions complémentaires : le LDSO pour une réponse locale d'échange de métadonnées intracommunautaire ; le RDSO pour une réponse globale d'échange de métadonnées interdisciplinaire et intercommunautaire. Cette section a pour objectif de valider expérimentalement les solutions que nous proposons.

Les questions à évaluer : Pour un échange de métadonnées intracommunautaire, nous souhaitons répondre à trois questions sur la proposition de Lac de Données de la Science Ouverte (LDSO) :

— Q1 : Est-ce que le LDSO permet un échange de métadonnées intracommunautaire ?
— Q2 : Est-il possible de permettre une proposition personnalisable pour les besoins et les connaissances des entités de recherche ?
— Q3 : Est-ce que le LDSO permet un gain de temps aux entités de recherche dans le processus de recherche de données de recherche ?

Pour une recherche unifiée, intercommunautaire et interdisciplinaire de données, nous souhaitons évaluer deux questions sur le Réseau de Données pour la Science Ouverte (RDSO) :

— Q4 : Est-ce que le RDSO permet un échange de métadonnées interdisciplinaire et intercommunautaire ?
— Q5 : Quelle est l'amélioration du niveau d'échange de métadonnées dans la Science Ouverte (observé dans la section 3.3.2) apporté par le RDSO ?

Pour répondre à ces questions, nous proposons deux expérimentations se basant sur le développement d'une preuve de concept (“Proof of Concept” - POC) de chaque solution et d'un retour d'expérience utilisateur. Pour chaque expérimentation, nous proposons une description des données utilisées, de l'environnement d'expérimentation et du protocole

expérimental. Nous présentons les résultats puis nous les analysons pour répondre aux questions posées précédemment. Dans la section 6.1, nous présentons l'expérimentation sur le LDSO permettant de répondre aux questions Q1, Q2 et Q3. Dans la section 6.2, une expérimentation sur le RDSO permet de répondre aux questions Q4 et Q5.

6.1 Evaluation du LDSO

Nous rappelons les trois questions de recherches sur le LDSO auxquelles nous souhaitons répondre :

— Q1 : Est-ce que le LDSO **permet un échange de métadonnées intracommunautaire** ?
— Q2 : Est-il possible de **personnaliser la proposition pour les besoins** et les connaissances des entités de recherche ?
— Q3 : Est-ce que le LDSO **offre un gain de temps aux entités de recherche** dans le processus de recherche de données de recherche ?

Pour répondre aux questions Q1 et Q2, nous avons réalisé une expérience utilisateur afin de comparer le processus de recherche de données sur le LDSO et sur trois PDRO existantes (AERIS [1], ODATIS [2], RCS-PDB [3]). Nous avons développé une preuve de concept du LDSO implémentant l'ensemble des fonctionnalités nécessaires à la recherche de données (cf. Figure 6.1). Cette preuve de concept se base sur le développement des outils d'accès et de gestion de métadonnées de la zone de gouvernance du LDSO.

Pour le choix des PDRO, nous avons choisi d'illustrer la recherche d'information réalisée par une entité de recherche de la communauté d'astrobiologie souhaitant mener un projet de recherche mêlant biologie, avec la base de données de protéines (RCS-PDB), et science environnementales, avec ODATIS (données océanographiques) et AERIS (données atmosphériques). Chaque plateforme utilise un modèle de métadonnées différent :

— ODATIS utilise une implantation de l'ISO 19115 ;
— AERIS utilise un modèle de métadonnée réalisé sur mesure pour les besoins applicatifs ;
— RCS-PDB utilise le modèle PDBx/mmCIF Dictionary.

6.1.1 Les données pour l'expérimentation

— 1 instance de métadonnées depuis ODATIS ;
— 21 instances de métadonnées depuis AERIS ;
— 1929 instances de métadonnées depuis RCS-PDB.

6.1.2 Environnement d'expérimentation

Nous avons développé une POC du LDSO. Cette preuve de concept intègre le développement des outils de la zone de gouvernance (accès et gestion des métadonnées). Elle implémente les éléments suivants (cf. Figure 6.1) :

— Une interface graphique web proposant un formulaire de requête des informations, adaptée aux besoins et aux connaissances des utilisateurs de notre expérimentation ;

1. `https://www.aeris-data.fr/`
2. `https://www.odatis-ocean.fr/`
3. https ://www.rcsb.org/

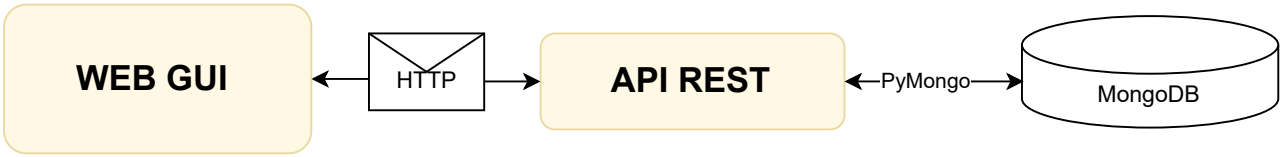

FIGURE 6.1 – Preuve de Concept du LDSO

— Une base de données NoSQL orientée document, avec MongoDB, intégrant la gestion multi-modèles avec les mappings ;
— Une API RESTful donnant accès au service de gestion de métadonnées, utilisant la gestion multi-modèles de la base de données MongoDB.

Cette interface graphique est ajoutée pour simplifier les interactions avec le LDSO. L'interface graphique web communique avec l'API RESTful par des requêtes HTTP. La connexion entre l'API RESTful et la base de métadonnées est réalisée via la bibliothèque PyMongo, fournissant un connecteur à la base de données MongoDB en Python. Tout le code et les données de cette expérimentation sont disponibles sur un dépôt Github public [4].

Pour suivre l'architecture de la base de métadonnées définie dans le chapitre 4, nous avons intégré dans une même collection ("metadata") les instances des métadonnées provenant des trois plateformes. Les mappings entre les différents concepts des modèles sont ingérés dans une même collection ("mapping"). Ces mappings sont effectués sur des concepts des modèles de métadonnées extraits des instances de métadonnées. Ces modèles sont conservés dans une même collection ("model").

Ces expérimentations se déroulent dans des conteneurs Docker, sur un ordinateur portable avec 16 Go de RAM et un processeur Intel(R) Core(TM) i5-1135G7. L'ensemble du code est implémenté en Python 3.8. Les analyses des résultats sont réalisées dans un notebook Jupyter. Le code de la POC et du scénario et les données sont disponibles sur un dépôt [5].

6.1.3 Protocole d'expérimentation

Notre expérimentation est réalisée avec la participation de 11 personnes. Ces personnes proviennent du domaine de la recherche et ont des connaissances en gestion de données. Nous leur avons demandé de réaliser huit requêtes sur les trois plateformes existantes puis sur la preuve de concept du LDSO permettant d'observer des informations différentes que nous détaillons après. Ces requêtes sont les suivantes :

— R1 : La liste des jeux de données publiés après le 01/01/2020 ;
— R2 : La liste des jeux de données associés à une publication contenant "Investigation" dans son titre ;
— R3 : La liste des jeux de données contenant des données sur la période du 01/01/2018 au 31/12/2022 ;
— R4 : La liste des jeux de données contenant des données sur une zone géogrpahique (Latitude nord : 90, latitude sud : -90, longitude est : 180, longitude west : -180) ;
— R5 : La liste des jeux de données associés à une publication publiée après l'année 2018 ;

4. `https://github.com/vincentnam/opendatalake_expe`
5. `https://github.com/vincentnam/opendatalake_expe`

— R6 : La liste des jeux de données associés à une publication dans un journal avec dans son titre “Env” ou “Meteo” ;
— R7 : La liste des jeux de données avec le mot-clé “oxido” ;
— R8 : La liste des jeux de données avec le mot-clé “oxido”, associé à une publication dans un journal avec un titre contenant “Env” ou “Meteo” ou sur une zone géographique (Latitude nord : 90, latitude sud : -90, longitude est : 180, longitude west : -180).

Une fois ces huit requêtes réalisées sur les 4 plateformes, nous avons posé six questions aux utilisateurs :

— Qe 1 : Êtes-vous : pas du tout confortable avec les plateformes de recherche de données de recherche (0 plateforme utilisée), un peu à l'aise avec les plateformes de recherche de jeux de données (quelques plateformes utilisées) ou à l'aise avec des plateformes de recherche de jeu de données (des plateformes régulièrement utilisées) ?
— Qe 2 : Combien de plateformes de recherche de jeux de données connaissez-vous ?
— Qe 3 : Connaissiez-vous les trois plateformes de l'expérimentation ?
— Qe 4 : Sans connaître ces plateformes, pensez-vous que vous auriez été capable de trouver les jeux de données que vous avez trouvés ?
— Qe 5 : De 1 à 5, à quel point jugez-vous nécessaire un accès unifié aux données ?
— Qe 6 : Laquelle des trois plateformes préférez-vous (AERIS, ODATIS, RCS PDB) et pourquoi ?

Ces requêtes ont été conçues pour rechercher des informations sur des concepts élémentaires (R1 (date de publication), R2 (titre de la publication), R5 (date de diffusion de la publication), R6(titre du journal de la publication), R7(mot clés)) et sur des concepts dépendant du type et du contenu de données (R3 (données temporelles), R4 (données géographiques)). Une dernière requête a été définie, comme une composition de requêtes des deux groupes de requête (R7, R6, R4).

L'objectif est de vérifier les capacités de recherche d'information du LDSO :

— sur des informations générales présentes dans la presque totalité des modèles ;
— sur des informations sur le contenu des données dépendant du type de données (spécifiques à certaines communautés ou disciplines) ;
— sur une recherche d'information complexe combinant les deux types d'informations.

L'ensemble de ces requêtes ne sont pas exécutable sur la totalité des plateformes (cf. Table 6.1). Ainsi, pour certains couples plateforme / requête, la réponse attendue est “il n'est pas possible de réaliser cette requête”. L'ensemble des requêtes est réalisable sur le LDSO.

Une interface de recherche d'information suffisamment claire et simple d'utilisation doit permettre aux utilisateurs de comprendre ce qui est possible et ce qui ne l'est pas avec les outils proposés. Cette possibilité de ne pas avoir de réponses à une requête nous permettra de confirmer si notre solution permet une adaptation des outils aux besoins et aux connaissances des utilisateurs cibles.

Les six questions que nous ajoutons permettent d'obtenir d'autres informations :

— **Population de l'étude :** La question Qe1 est destinée à permettre une stratification de la population d'étude. L'étude du SCNAT en Suisse sur les données ouvertes[6] montre une répartition de la population en fonction de son savoir sur la manière d'ouvrir les données (cf. Q12 de l'étude du SCNAT) : $\sim$ 40% savent comment ouvrir leurs données, $\sim$ 40% ont quelques connaissances sur l'ouverture

6. https://map.scnat.ch/en/activities/open_data_survey

	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 6.1 – Les possibilité de requête sur les plateformes
X : La requête est réalisable sur la plateforme ; * : Les requêtes de géolocalisation sur ODATIS sont réalisées via une carte. Ce type d'outil ne permet pas un requêtage précis. **PDB ne permet de rechercher des titres complets ou des chaînes de caractères au sein d'un titre

des données et $\sim$ 20% ne savent pas du tout. Ces 11 utilisateurs sont catégorisés selon la question Qe1.

— **Etat des connaissances de la Science Ouverte :** La question Qe2, Qe3, Qe4 servent à apprécier le niveau de connaissance des personnes participant à l'étude afin de s'assurer que cette population ressemble à l'entité de recherche ER du scénario servant de fil rouge à cette thèse. **Confirmation des besoins :** La question Qe 5 permet de confirmer que le besoin d'accès unifié aux données, exprimé dans la littérature, est aussi un besoin exprimé dans notre population d'étude. La Question Qe6 permet d'apprécier les besoins exprimés par les utilisateurs, en termes d'outils de recherche de données.

6.1.4 Évaluation

Nous avons trois objectifs d'évaluation dans cette expérimentation : la capacité d'échange intracommunautaire du LDSO, la réponse aux besoins des entités de recherche, le gain de temps apporté par le LDSO.

Résultats : Notre population est répartie comme suit selon la question Qe1 :

— 4 utilisateurs n'ayant aucune connaissance sur l'utilisation de plateformes de gestion de données ($\sim$ 36%) ;
— 4 utilisateurs ayant quelques connaissances sur l'utilisation ($\sim$ 36%) ;
— 3 utilisateurs étant à l'aise sur l'utilisation des plateformes de recherche de jeux de données ($\sim$ 27%).

Pour évaluer le gain de temps apporté par le LDSO, nous avons compilé les temps de requis pour obtenir les résultats des requêtes des utilisateurs sur les différentes plateformes. Nous avons extrait le temps moyen pour une requête sur chacune des plateformes, en considérant la prise en compte ou non des requêtes qui ont abouti à des erreurs (cf. Tableau 6.2). La plateforme ayant le plus long temps de requêtage est la plateforme RCSB PDB. Le temps de requêtage avec erreur montre une moyenne d'environ $\sim$ 10% plus élevée que le temps sans les erreurs pour RCS PDB.

Temps moyen de requête (en secondes)	AERIS	ODATIS	RCSB PDB	Total	LDSO
Avec erreur	26.74	22.73	31.08	80.55	22.96
Sans erreur	27.84	21.67	34.32	83.83	22.93

TABLE 6.2 – Temps moyen de requête par plateforme

Pour la réponse aux besoins des utilisateurs, nous avons compilé le nombre d'erreurs

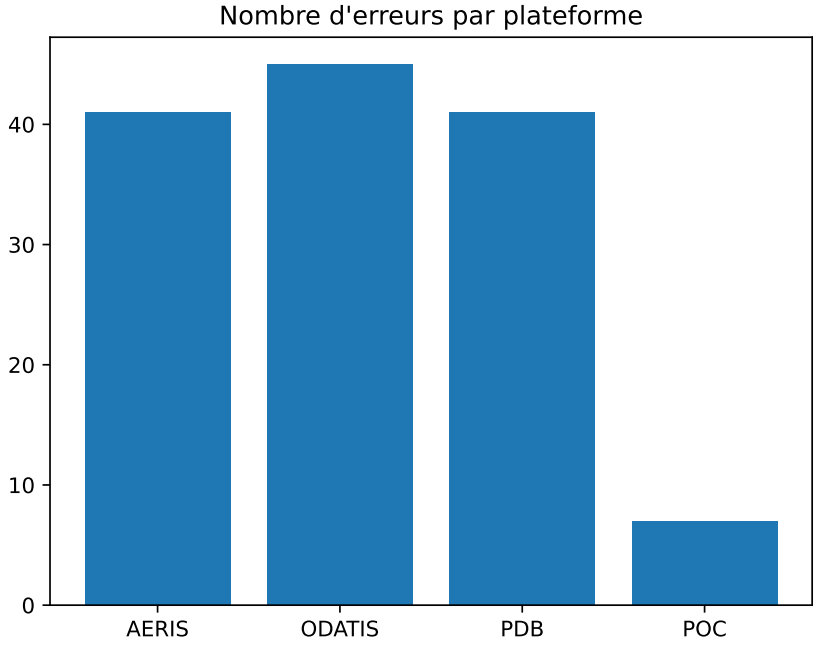

FIGURE 6.2 – Nombre d'erreurs par plateforme, pour un total de 88 requêtes effectuées par plateforme

des utilisateurs dans les requêtes par plateformes (cf Tableau 6.2). Nous observons que le LDSO engendre près de quatre fois moins d'erreurs dans le requêtage que les autres plateformes. En effet, nous avons développé cette solution pour s'adapter aux connaissances des utilisateurs et utiliser un vocabulaire auquel ils sont habitués. Nous montrons ainsi que cette solution peut être personnalisable et répondre au plus près des besoins des utilisateurs.

Pour les réponses aux questions (en dehors de la Q1 qui a servi à réaliser l'échantillonnage), les utilisateurs ont répondu :

— Qe 2 : Les 4 utilisateurs ($\sim$ 36%) n'étant pas à l'aise avec les solutions de recherche de données de recherche ont indiqué qu'elles connaissaient 0 plateformes, sauf un utilisateur connaissant 1 plateforme. Les 7 autres utilisateurs ($\sim$ 64%), un peu à l'aise ou à l'aise avec les plateformes de recherche de jeux de données, ont indiqué qu'ils connaissaient 4 ou 5 plateformes, sauf 2 utilisateurs en connaissant 2 ou 3.

— Qe 3 - Qe 4 : Tous les utilisateurs (11) ont répondu que les trois plateformes de l'expérimentation étaient inconnues. 5 utilisateurs ($\sim$ 45%) ont répondu qu'ils ne pourraient pas ou peut-être pas retrouver les données. 6 utilisateurs ($\sim$ 55%) ont répondu qu'ils arriveraient ou qu'ils arriveraient peut-être à retrouver ces données. Nous notons que 5 utilisateurs ayant répondu "Oui" à la question Qe4 ont indiqué que cette démarche leur demanderait du temps.

— Qe 5 : La totalité des utilisateurs a répondu qu'un accès unifié aux données était nécessaire (4) ou extrêmement nécessaire (5).

— Qe 6 : Les utilisateurs forment deux groupes : 5 utilisateurs ($\sim$ 45%) ont préféré la plateforme PDB pour la richesse apportée par le système de requêtage, 6 utilisateurs ($\sim$ 55%) ont préféré la plateforme AERIS pour la simplicité d'utilisation du système de recherche.

Analyse des résultats : Nous analysons ces résultats avec quatre focalisations différentes :

— la population de l'étude ;
— le gain de temps apporté par le LDSO ;
— la réduction du nombre d'erreurs par le LDSO dans les requêtes ;
— l'état de connaissance de la Science Ouverte des utilisateurs.

Population de l'étude : Notre distribution possède une surreprésentation des utilisateurs n'ayant aucune connaissance dans l'utilisation des plateformes de recherche de jeu de données tandis que nous avons une sous-représentation des utilisateurs étant à l'aise avec ces plateformes. La question Qe 5 permet de confirmer que le besoin d'accès unifié aux données, exprimé dans la littérature, est aussi un besoin exprimé dans notre population d'étude. La Question Qe6 permet d'apprécier les besoins exprimés par les utilisateurs, en terme d'outils de recherche de données. L'utilisation de plateformes de recherche est une étape plus avancée dans le processus d'adoption de la Science Ouverte. De plus, les besoins exprimés par les utilisateurs sont les mêmes que ceux exprimés dans la littérature. Malgré la différence observée entre notre population et la population ciblée, nous supposons que notre distribution s'approche de la distribution réelle.

Gain de temps apporté par le LDSO : Nous expliquons la différence observée entre les requêtes avec et sans erreur sur la plateforme RCS PDB par la richesse du système de requêtage proposé. Nous observons ensuite que le LDSO fournit un outil ayant un temps similaire à celui des autres plateformes. Cependant, les requêtes réalisées sur le LDSO permettent de réaliser des requêtes sur les 4 plateformes en même temps, réduisant le temps de requêtage pour un chercheur à $\frac{1}{nb}$, où nb est le nombre de plateformes intégrées en comptant le LDSO.

Réduction du nombre d'erreurs par la personnalisation des outils du LDSO : Les utilisateurs ont fait un nombre d'erreurs équivalent entre les trois PDRO sélectionnées. Cependant, le nombre d'erreurs réalisées sur la POC du LDSO est en moyenne ~ 4 fois inférieur aux autres plateformes. Ces résultats indiquent que notre solution offre bien des capacités de personnalisation des outils pour s'adapter au mieux aux utilisateurs et à leurs besoins. Cette personnalisation nous permet de fournir une solution compréhensible et facilement utilisable.

Etat des connaissances de la Science Ouverte : Les réponses apportées à la question Qe 2 nous confirment que la connaissance des plateformes de gestion de données de la Science Ouverte dans notre population est très réduite, avec un maximum de cinq plateformes. Cette information permet de confirmer que le nombre de plateformes connues par l'entité de recherche ER est fidèle aux populations cibles et donc de valider le profil de l'entité de recherche ER. Les résultats de Qe 3 et la Qe4 nous indiquent qu'une recherche d'information sur des plateformes inconnues des utilisateurs permettrait à près de 45% d'entre eux de trouver des informations qu'ils ne pourraient trouver autrement. Nous observons un manque de connaissance des solutions de gestion de données ouvertes qui pose un problème important pour la recherche d'informations. La réutilisation des informations apportées par chaque utilisateur au sein du LDSO, à travers l'intégration des informations de plateformes externes, est profitable à l'ensemble des utilisateurs du LDSO.

6.1.5 Bilan

Ces résultats nous permettent donc de conclure sur les questions de recherche Q1, Q2 et Q3 sur le LDSO :

— Q1 : La mise en place d'une preuve de concept en intégrant plusieurs plateformes différentes utilisant des modèles différents montre la capacité du LDSO à mettre en place un échange d'information intracommunautaire. De plus, notre solution permet de proposer une plus grande richesse de requêtage sur les métadonnées que les plateformes évaluées dans l'étude (cf. Tableau 6.1).

— Q2 : Nous avons montré que la solution du LDSO est personnalisable et adaptable aux besoins et aux connaissances des utilisateurs car le nombre d'erreurs est plus faible à l'aide de notre solution. De plus, la proposition d'un accès unifié aux données est un besoin très présent au sein des communautés de chercheurs (cf. Qe5). Le LDSO permet de fournir cet accès unifié aux données et donc de répondre à ce besoin des chercheurs.

— Q3 : Nous avons montré que le temps de recherche sur le LDSO est similaire à celui nécessaire aux autres solutions, mais permet d'aller chercher des informations provenant de plusieurs plateformes avec une seule requête. Le LDSO permet de réduire le temps de recherche de données inversement proportionnellement au nombre de plateformes intégrées dans le LDSO.

De plus nous avons observé des informations supplémentaires à nos questions de recherche :

— Les résultats indiquent que les connaissances de la Science Ouverte de l'entité de recherche ER sont proches de la réalité.

— Le manque de connaissance sur l'existence des PDRO est un frein majeur à la réutilisation de données.

Nous avons réussi à valider l'ensemble de nos questions de recherche et proposer une solution permettant un échange de métadonnées intercommunautaire et une recherche de données unifiée, qui s'adapte aux besoins des utilisateurs et leur permet de gagner du temps sur le processus de recherche d'information.

6.2 Evaluation du RDSO

Le RDSO est une solution dont l'objectif est l'échange interdisciplinaire et intercommunautaire de métadonnées et la recherche transparente interdisciplinaire et intercommunautaire. Pour valider cette solution, nous avons défini deux questions auxquelles nous souhaitons répondre par expérimentation, que nous rappelons :

— Q4 : Est-ce que le RDSO **permet un échange de métadonnées interdisciplinaire et intercommunautaire** ?

— Q5 : Quelle est l'**amélioration du niveau d'échange de métadonnées** dans la Science Ouverte (observé dans la section 3.3.2) apporté par le RDSO ?

Pour répondre aux questions Q4 et Q5, nous avons élaboré un scénario d'expérimentation avec une preuve de concept du RDSO.

6.2.1 Les données pour l'expérimentation

Avant de présenter les données utilisées, nous souhaitons valider que notre solution peut être appliquée à la totalité des modèles de métadonnées de la Science Ouverte. Cette

évaluation permet de s'assurer que la sélection de nos données n'apporte aucun biais.

Ainsi, nous avons téléchargé 17 modèles de métadonnées en suivant la catégorisation proposée par Ulrich et al. (2022), selon leur type (technique, sémantique), le domaine ou le cas d'utilisation. Pour étudier la différence dans les choix de modélisation selon les communautés ou les acteurs, nous avons sélectionné des organismes de standardisation différents (cf. Tableau 6.3).

Domaine \ Cas d'utilisation	Recherche d'informations	Intégration des données	Ensemble de données de base	Réutilisation
Humanités et Sciences Sociales	**(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
Sciences de la Vie	(Sem) (WHO) ICD-10	**(3)** : (Struct) (TDWG) Darwin Core		**(8)** : (Struct) (HL7) FHIR **(2)** : (Struct) (HL7) C-CDA
Sciences Naturelles	**(12)** : (Struct) PDB		**(9)** : (Struct) (ISO) ISO-19115 **(1)** : (Struct) AERIS	
Sciences de l'Ingénieurie	**(7)** : (Struct) (RDA) EngMeta – Metadata for Computational Engineering		**(15)** : (Struct) (OGC) SensorML **(17)** : (Struct) (OGC) CoverageJSON	
Général	(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 6.3 – Classification des modèles de métadonnées

Nous avons effectué manuellement des mappings entre ces différents modèles (cf. Tableau 6.4). Nous avons sélectionné 3 concepts différents : le titre du jeu de données, la localisation du contenu et l'unité de mesure.

Concept \ Modèle (cf. Tableau 6.3)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Titre du jeu de données	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Localisation du contenu	x		x					x		x							x
Unité de mesure	x	x					x								x		

TABLE 6.4 – Interopérabilité des modèles : les concepts alignés dans ces modèles

Nous avons pu mettre en place une interopérabilité de l'ensemble des modèles. La Figure 6.3 illustre le graphe des interopérations entre modèles. Nous observons que ce graphe est proche d'un graphe dense (avec une densité de 0.91). L'interopération de l'ensemble des modèles de métadonnées montre la capacité d'intégration de ces modèles provenant de la Science Ouverte dans le RDSO.

Maintenant que nous avons validé que notre solution peut intégrer n'importe quel modèle de métadonnées de la Science Ouverte, nous présentons la sélection de données pour notre expérimentation.

Nous avons sélectionné quatorze plateformes en s'assurant qu'elles possèdent les caracctéristiques suivantes (cf. Figure 6.4) :

— des quatre grandes disciplines de la recherche (Humanité et Sciences Sociales (en jaune dans la Figure 6.4), des Science Naturelles (en vert dans la Figure 6.4), des Sciences de l'Ingénieur (en bleu dans la Figure 6.4) et des Sciences de la Vie (en rouge dans la Figure 6.4). Nous avons ajouté des plateformes générales, pouvant gérer des données de toutes les disciplines (en gris dans la Figure 6.4) ;
— des plateformes gérant peu de jeux de données (quelques dizaines) comme de très nombreux jeux de données (plusieurs milliards) ;
— des plateformes provenant de communautés différentes, comme des plateformes européennes et des plateformes américaines.

Ces plateformes sont les suivantes :

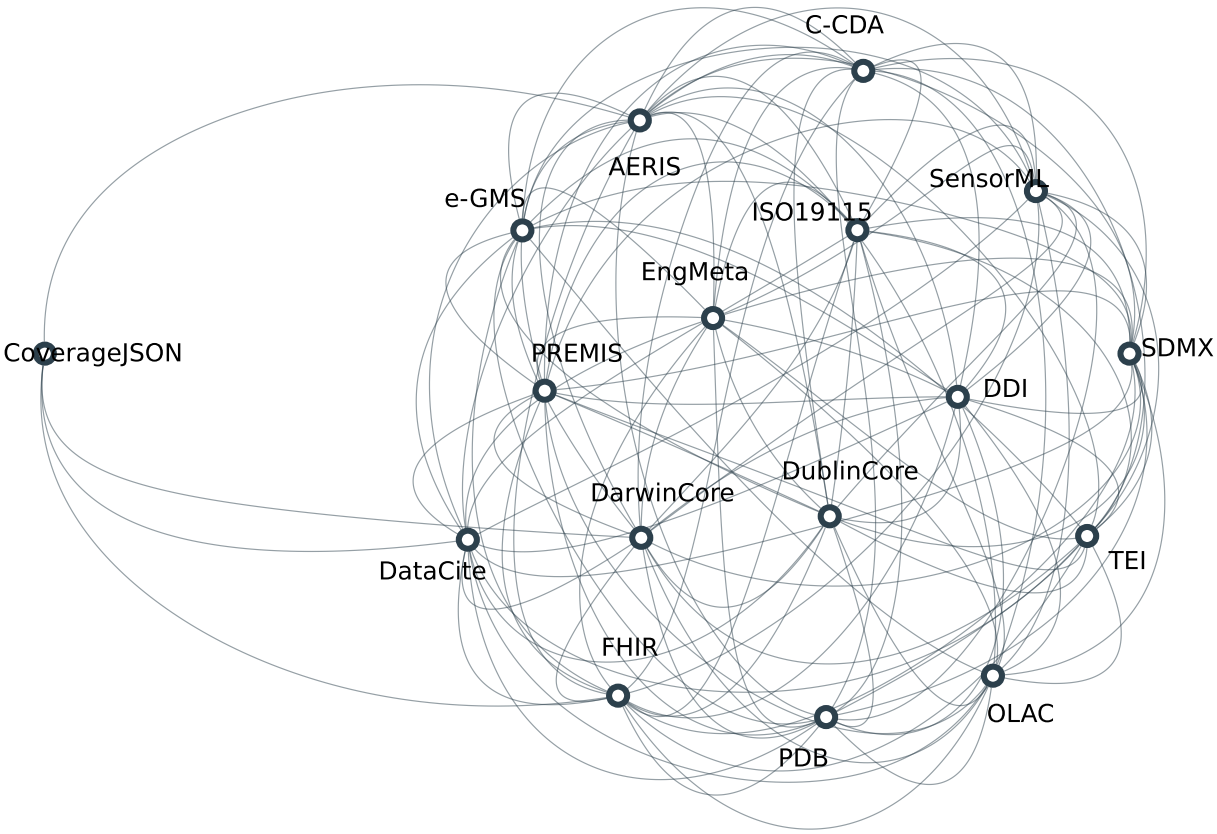

FIGURE 6.3 – Illustration des interopérations des modèles

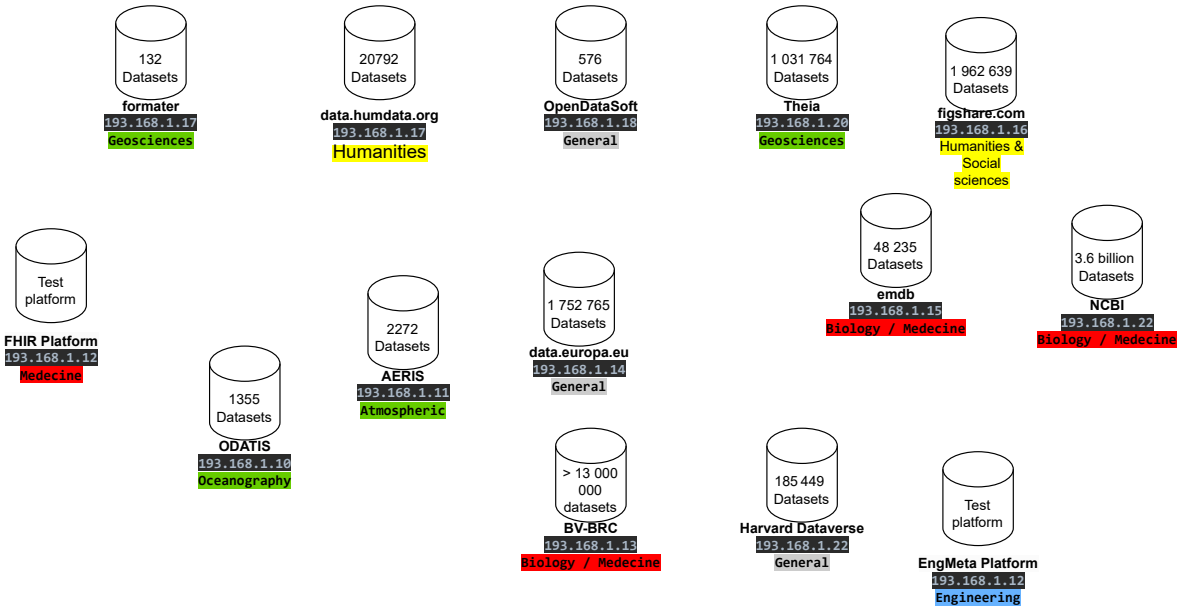

FIGURE 6.4 – Les plateformes sélectionnées pour les expérimentations (au 3 Juillet 2024)

— Formater [7] : nous avons téléchargé 3 instances de métadonnées aléatoires (en format

7. https://www.poleterresolide.fr/

JSON).
- The Humanitarian Data Exchange (data.humdata.org) : nous avons téléchargé 8 instances de métadonnées (en format JSON) en lien avec le travail ("job") ou le virus du Covid19.
- OpenDataSoft [8] : nous avons téléchargé le catalogue entier (576 instances de métadonnées, au format CSV).
- Theia [9] : nous avons téléchargé 6 instances de métadonnées aléatoires (au format JSON).
- Figshare [10] : nous avons téléchargé 13 instances de métadonnées en lien avec le Covid ou des outils de biocontrôle (au format XML)
- ODATIS [11] : nous avons téléchargé 1 instance de métadonnées au hasard (au format XML)
- AERIS [12] : nous avons téléchargé 21 instances de métadonnées aléatoires
- Data Europa [13], la plateforme de données ouvertes de l'Union Européenne : nous avons téléchargé 8 instances de métadonnées autour des outils de biocontrôle, des données sociales dans des pays européens (au format JSON)
- EMDB [14] : nous avons téléchargé 2 instances de métadonnées autour du virus du COVID 19 (au format XML)
- NCBI [15] : nous avons téléchargé 18 instances de métadonnées autour des outils de biocontrôle (au format CSV)
- BV-BRC [16] : nous avons téléchargé 22 instances de métadonnées aléatoires (au format CSV)
- Harvard Dataverse [17] : nous avons téléchargé 10 instances de métadonnées autour des outils de biocontrôle, du virus du COVID 19 et des données sociologiques (au format CSV)

Nous avons aussi ajouté deux plateformes simulées (les plateformes "Test platform" dans la Figure 6.4) : une plateforme utilisant le modèle FHIR pour simuler l'intégration d'un organisme de soin ; une plateforme utilisant le modèle EngMeta pour simuler l'intégration d'une plateforme de gestion de données provenant du domaine de l'ingénierie.

6.2.2 Environnement d'expérimentation

Nous avons développé une POC du module du RDSO. Nous avons ensuite déployé une API REST et un système de gestion de métadonnées pour chaque plateforme de cette POC.

Nous avons déployé un RDSO sans aucune plateforme, et nous avons à tour de rôle inscrit les plateformes dans ce RDSO d'expérimentation en utilisant le protocole d'inscription du RDSO (cf. Figure 6.5). Pour toutes ces plateformes, nous avons enrichi le registre avec les informations :

8. https://www.opendatasoft.com/fr/
9. https://www.theia-land.fr/
10. https://figshare.com/
11. https://www.odatis-ocean.fr/
12. https://www.aeris-data.fr/
13. https://data.europa.eu/
14. https://www.ebi.ac.uk/emdb/
15. https://www.ncbi.nlm.nih.gov/
16. https://www.bv-brc.org/
17. https://dataverse.harvard.edu/

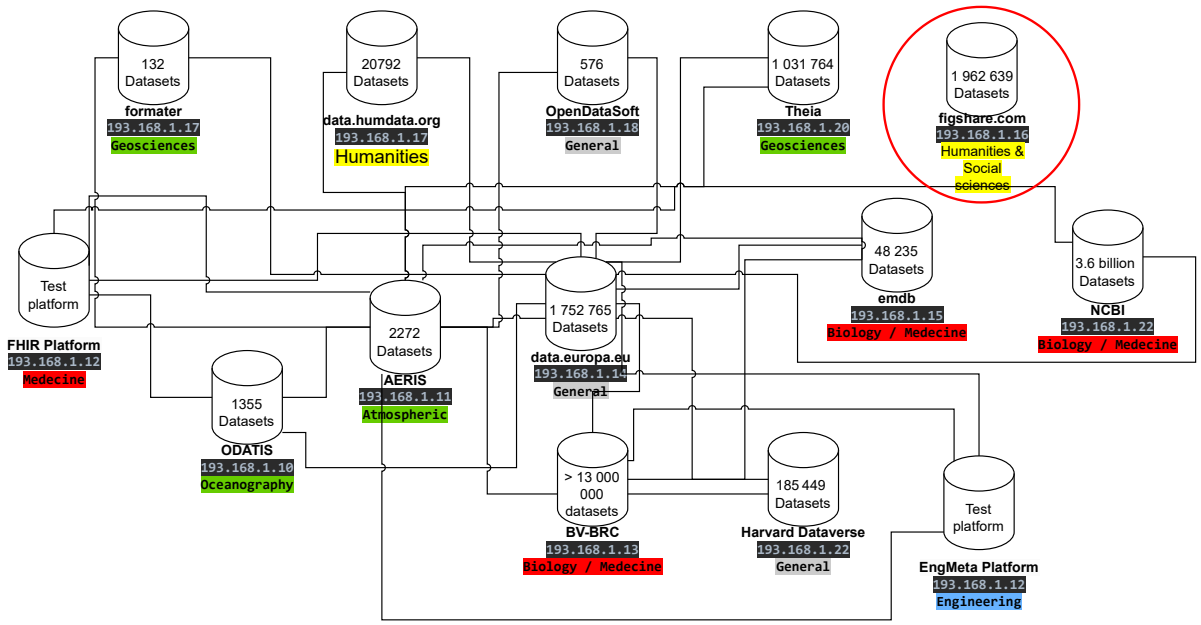

FIGURE 6.5 – Etat initial du RDSO dans notre expérimentation

— de toutes les plateformes ;
— de tous leurs modèles de métadonnées.

Ensuite, nous avons mis en place 54 mappings à la main entre les concepts des différents modèles des plateformes que nous avons insérés dans le registre (cf. la définition du registre contenant ces mappings [18]).

Ces expérimentations se déroulent dans des conteneurs Docker, sur un ordinateur portable avec 16 Go de RAM et un processeur Intel(R) Core(TM) i5-1135G7. L'ensemble du code est effectué en Python 3.8. Les analyses des résultats sont réalisées dans un notebook Jupyter. L'ensemble du code (POC et expérimentation), ainsi que les données et des scripts permettant de réexceuter la totalité de l'expérimentation sont disponibles sur un dépôt Github public [19].

6.2.3 Protocole d'expérimentation

Sur la base de ce RDSO initial, nous avons conçu un scénario de validation de ses différentes fonctionnalités. Le scénario d'expérimentation se déroule en trois étapes :

— L'inscription d'une nouvelle plateforme, permettant de vérifier le bon fonctionnement des mécanismes de propagation des modifications du registre ;
— L'interopération du modèle de métadonnées de cette plateforme dans le registre, permettant de vérifier le bon fonctionnement de la gestion des modèles de métadonnées avec les mappings ;
— Des exemples de requêtes montrant l'enrichissement de la quantité de réponses grâce à la propagation des requêtes au sein du RDSO.

Pour réaliser ce scénario, nous avons utilisé la plateforme Figshare comme nouvelle plateforme s'inscrivant dans la plateforme. Nous avons inséré les informations de plate-

18. https://github.com/vincentnam/Openscience_network_experiment/blob/master/experiments/Part1-API/registry.json
19. https://github.com/vincentnam/Openscience_network_experiment

forme, de modèle de métadonnées de la plateforme Figshare, et nous avons défini deux mappings avec ce modèle. Une vidéo démonstration de toute cette étape est présente dans le dépôt Github.

Ensuite, pour répondre aux questions de recherche Q4 et Q5, nous avons demandé à un chercheur (Dr. Pressecq (OrcID 0000-0003-0067-7903), chercheur en agronomie) de réaliser les requêtes qu'il a effectuées lors de son projet de recherche afin d'observer les apports du RDSO dans cette situation.

Son projet de recherche a visé à développer des outils d'aide à la décision pour l'utilisation d'outils de biocontrôle pour les agriculteurs (Pressecq (2023)).

Durant les trois ans de son projet de recherche, le Dr Pressecq a dû chercher des données de recherche (principalement des articles scientifiques) sur différents agents microbiens. Les requêtes qu'il a réalisées sont simples, permettant de chercher tous les jeux de données qui contiennent dans le titre ou la description le nom de la souche. Nous lui avons demandé de réaliser les mêmes requêtes sur notre POC de RDSO, sur une souche particulière (à savoir "Trichoderma Harzanium T-22").

6.2.4 Évaluation

Résultats : Dr. Pressecq a conservé 17 jeux de données grâce au RDSO. Ces jeux de données proviennent de 3 plateformes différentes (13 résultats sur le NCBI, 3 depuis Harvard Dataverse et 2 sur la plateforme Figshare). Sur ces 17 résultats, 12 étaient de nouveaux jeux de données initialement inconnus du Dr. Pressecq. Le temps total passé à la recherche d'informations est de 2 heures et 30 minutes.

Analyse des résultats : Nous analysons ces résultats selon quatre focalisations différentes :

— l'enrichissement du volume de données ;
— le gain de temps apporté ;
— la capacité de rechercher des données intercommunautaires et interdisciplinaires ;
— l'enrichissement du processus de recherche scientifique lié à cette recherche intercommunautaire.

Le projet de recherche du Dr. Pressecq l'a conduit à extraire des informations provenant de 900 publications scientifiques sur 41 souches de micro-organismes différents, lui permettant de construire un jeu de données sous forme de tableau contenant 381 lignes. Pour son projet de recherche, il a utilisé cinq plateformes différentes : deux pour la recherche d'articles scientifiques (Google Scholar, Web Of Science) où il a trouvé la majorité de ses données, et trois permettant de chercher des jeux de données (la plateforme de données ouverte du gouvernement Français [20], la plateforme de gestion de données agricole et agroalimentaire AgDataHub [21] et Smartbiocontrol [22]).

La partie de son jeu de données sur cette souche particulière a nécessité d'extraire des informations de 115 articles scientifiques pour générer 29 lignes de son jeu de données. Il a estimé le temps moyen pour générer une ligne à 79 minutes, pour un total d'un peu moins de 56h de travail pour 29 lignes de son jeu de données. Chaque ligne de son jeu de données contient une combinaison entre un agent microbien et ses facteurs d'efficacité associés (propriété de l'agent de biocontrôle, conditions environnementales, pratiques de cultures, propriétés du pathogène).

20. `data.gouv.fr`
21. `https://agdatahub.eu/`
22. `http://www.smartbiocontrol.eu/`

Enrichissement du volume de données par le RDSO : Les résultats de recherche ont permis au Dr. Pressecq de trouver de nouvelles données qu'il n'avait pas pu trouver durant son projet de recherche. Les résultats de recherche d'information sur le RDSO lui ont permis d'ajouter 2 nouvelles lignes à son jeu de données. Cela correspond à un **enrichissement du volume de son jeu de données de 7%**, permis par le RDSO.

Gain de temps apporté par le RDSO : La construction de son jeu de données a nécessité d'extraire des informations depuis de nombreux articles scientifiques. Dans les résultats de recherche, trois lignes de son jeu de données auraient pu être réutilisées depuis un jeu de données qu'il n'avait pas trouvé. Trouver ces trois lignes lui a demandé un total de 45 minutes. Pour construire ces lignes, après extraction d'information depuis des articles scientifiques lui ont demandé 237 minutes. Cela représente une **réduction du temps** de plus **de 80%**, permise grâce à la réutilisation de données apportées par le RDSO.

Recherche de données intercommunautaire et interdisciplinaire du RDSO : Les plateformes d'où proviennent les jeux de données sont le NCBI, Harvard Dataverse et la plateforme Figshare. Cette recherche a permis de mener une recherche de données sur une plateforme provenant du domaine de la biologie et de la médecine, une plateforme généraliste et une plateforme provenant du domaine des humanités et des sciences sociales. Ces résultats montrent la **mise en place claire d'un échange d'information intercommunautaire et interdisciplinaire**.

Enrichissement du processus de recherche scientifique par le RDSO : Le Dr. Pressecq a trouvé des jeux de données lui permettant de répondre à d'autres questions de recherche. Ces jeux de données n'ont pas été inclus dans les 17 jeux de données trouvés lors de cette expérimentation, car ils ne répondaient pas à la question de recherche actuelle. Le Dr. Pressecq a trouvé des jeux de données provenant de la plateforme data.europa.eu dans le domaine des Sciences Sociales, lui permettant de concevoir un nouveau projet de recherche interdisciplinaire, et des jeux de données provenant de la plateforme FORMATER lui permettant d'enrichir son analyse avec un croisement des données de caractéristiques des sols, lui permettant de réaliser une inspection de différents agents microbiens selon le type des sols.

Ces résultats nous ont donc permis d'observer trois impacts du RDSO sur la recherche d'information dans la Science Ouverte :

- Impact temporel : Le RDSO permet une réduction du temps nécessaire pour trouver des données important - plus de 80% dans notre expérimentation
- Impact volumétrique : Le RDSO permet aux chercheurs de trouver des jeux de données qu'ils n'auraient pas trouvés initialement, avec un enrichissement du volume du jeu de données de **7%** dans notre expérimentation
- Impact interdisciplinaire : Le RDSO fournit un accès transparent aux jeux de données de l'ensemble des plateformes inscrites dans le RDSO. Grâce à cet aspect, le Dr. Pressecq a réussi à trouver des données de plateformes qu'il ne connaissait pas et hors de son domaine de connaissances. Le RDSO lui a permis de trouver des jeux de données qui enrichissent ses questions de recherche mais aussi développent de nouvelles questions de recherche :
  - Des questions de **recherche intradisciplinaire**, en croisant ses données aux données de caractéristiques des sols trouvés sur la plateforme FORM@TER

pour une prospection de différents agents microbiens selon les types de sols ;
— Des questions de **recherche interdisciplinaire**, en croisant les données trouvées sur la plateforme européenne de recherche de données, afin d'évaluer la perception des outils de biocontrôle chez les agriculteurs et les consommateurs, permettant la mise en place d'une collaboration entre la communauté des chercheurs en agronomie et la communauté de chercheurs en sciences sociales.

Ces résultats de recherche ont nécessité de rendre interopérables les modèles avec des mappings faits à la main. Cependant, sur les 1136 concepts différents définis dans l'ensemble des modèles de métadonnées que nous avons intégrés dans notre scénario, les 56 mappings réalisés (les cinquante-quatre initiaux plus les deux réalisés lors de l'inscription de la plateforme figshare) ont été effectués sur 67 concepts différents. Le niveau d'interopérabilité des modèles de métadonnées est donc très faible.Le RDSO offre donc un apport positif à la recherche d'information malgré un faible niveau d'interopérabilité des modèles de métadonnées.

Donc, les apports du RDSO sont déjà significativement visibles malgré un très faible niveau d'interopération entre les modèles. Une seule personne a pu inscrire 14 plateformes au sein du RDSO. Nous avons observé que le coût de l'enrichissement du registre est faible.

Maintenant que nous avons validé la viabilité de notre proposition, nous souhaitons aller plus loin dans nos analyses en étudiant l'apport du RDSO à l'échange de métadonnées que nous avons évalué dans le chapitre 3. De plus, nous souhaitons mesurer l'apport du RDSO pour réduire le coût d'adoption de notre proposition.

6.2.5 Analyse du RDSO

Pour observer l'apport du RDSO dans la Science Ouverte, nous souhaitons évaluer deux caractéristiques du RDSO :

— Le niveau d'échange d'information après l'implantation du RDSO par rapport au niveau observé dans le chapitre 3 ;
— Le coût d'adoption de la solution, que les chercheurs jugent nécessaire de prendre en compte (Aydinoglu et al. (2014)).

Sur ces deux aspects, nous proposons une analyse comparative, avec et sans le RDSO, afin d'en mesurer quantitativement l'apport.

6.2.5.1 Le niveau d'échange de métadonnées

Nous avons évalué le niveau d'échange de métadonnées dans la Science Ouverte (cf. Section 3.3.2.3) compris dans l'intervalle

$$Pr(X_d) \in [\frac{1+8.89}{3117}; \frac{1+19.59}{3117}] = [0.003; 0.007]$$

avec un niveau de confiance à 99.9997% (cf. chapitre 3).

Pour évaluer l'impact du RDSO dans la Science Ouverte, nous avons compté le nombre de plateformes que nous arrivons à intéropérer avec les modèles de métadonnées que nous avons sélectionnés. Les six modèles de ces expérimentations (SDMX, Darwin Core, Dublin Core, DDI, ISO 19115, DataCite) sont utilisés par 72% des plateformes présentes dans les données de Re3Data. Avec notre solution, il y a un nombre de sauts infini pour aller chercher les données chez les voisins, à l'inverse du protocole OAI-PMH qui est limité à 1 (protocole utilisé pour évaluer l'état de l'échange de métadonnées actuel dans le chapitre 3). Notre solution permet d'interroger n'importe quelle plateforme du réseau

depuis n'importe quelle autre plateforme. Ainsi, la probabilité de trouver une donnée est égale à la proportion de plateformes que nous interconnectons.

$$Pr_{\infty}(X) = 0.72$$

Basé sur nos expérimentations, l'échange d'informations dans la Science Ouverte passerait de 0.7% (cf. Chapitre 3) à 72%, pour une multiplication par plus de 100 du niveau d'échange d'information. Or l'ensemble des modèles de métadonnées de la Science Ouverte peut être intégré dans le RDSO. Ainsi, en supposant une adoption par la totalité des plateformes, le niveau d'échange d'information théorique apporté par le RDSO est de 100%. Un niveau d'échange d'information effectif de 72% nous permet de rendre l'échange de métadonnées possible, et de déployer les étapes suivantes pour l'implantation de la Science Ouverte (Nosek (2019)).

Pour s'assurer de la viabilité de l'adoption du RDSO à l'ensemble de la Science Ouverte, il est nécessaire d'évaluer le coût d'adoption du RDSO.

6.2.5.2 Le coût d'adoption du RDSO

Pour évaluer le coût d'adoption du RDSO, nous comparons le RDSO à une interopération non organisée qui se ferait pair à pair entre deux plateformes qui se connaissent. Dans les hypothèses que nous prenons, nous évaluons le meilleur des cas pour la solution d'échange de métadonnées pair à pair pour nous assurer que l'évaluation soit la plus fiable possible.

Assurer l'interopérabilité de deux PDRO nécessite trois opérations : (1) l'implantation d'une API permettant l'échange d'informations avec une autre plateforme, (2) l'échange d'informations sur les modèles de métadonnées entre les deux plateformes et (3) l'interopération des modèles de métadonnées.

Adoption d'un échange de métadonnées pair à pair Notons C_1, C_2 et C_3, respectivement le coût des opérations (1), (2), et (3) pour mettre en place l'échange d'information pair à pair non organisé globalement. Supposons que nous avons n plateformes et m modèles.

Pour l'opération (1), chaque plateforme doit mettre en place une API compatible avec la plateforme à interconnecter. Dans le meilleur des cas, toutes les API possèdent la même technologie et sont par défaut compatibles entre elles. Chaque plateforme doit donc déployer une API. Le coût total de mise en place de l'opération (1) est égal à $n * C_1$, avec n le nombre de plateformes.

Pour l'opération (2), chaque plateforme doit envoyer ses informations à l'ensemble des autres plateformes. L'envoi d'information étant un processus unidirectionnel, deux envois sont nécessaires pour chaque couple de plateformes que nous souhaitons rendre interopérables. Le coût de l'opération (2) est donc égal à $2 * \binom{n}{2} * C_2$, avec n le nombre de plateformes.

Pour l'opération (3), chaque plateforme doit réaliser l'interopération de son modèle avec l'ensemble des autres modèles de métadonnées. Cette interopération pair à pair ne permet pas de réutiliser les mappings réalisés par les autres plateformes, par absence d'organisation de l'échange de ces informations. Le coût de l'opération (3) est $n * \binom{m}{2} * C_3$, avec m le nombre de modèle et n le nombre de plateformes.

Ainsi, le coût total de l'adoption d'une solution pair à pair non organisée globalement pour l'échange d'information au sein d'un groupe de n plateformes utilisant m modèles de métadonnées (avec une variété des modèles de métadonnées relativement faible) est égal à $C_{pap}(n, m) = n * C_1 + 2 * \binom{n}{2} * C_2 + n * \binom{m}{2} * C_3$, dans le meilleur des cas.

Le coût d'adoption du RDSO Dans le cas du RDSO, les mêmes opérations doivent être mises en place. Nous définissons les coûts C'_1,C'_2 et C'_3, respectivement le coût des opérations (1), (2) et (3) dans le cas de l'adoption du RDSO.

La POC du RDSO est réalisé via des conteneurs permettant un déploiement automatique et une réutilisation sans modification de la majeure partie du code du module. Le seul besoin dans ce déploiement est la mise en place d'une fonction d'interopération entre le module et les systèmes de gestion de métadonnées de la plateforme. Nous en déduisons que $C_1 \leq C'_1$. Chaque plateforme doit implanter le module du RDSO. Le coût de l'opération (1) est donc égal à $n * C'_1$.

La gestion de la propagation des informations est réalisée via le module du RDSO et le mécanisme de propagation des informations est intégré au module. Ainsi, chaque plateforme n'a à renseigner qu'une seule fois ses informations qui sont automatiquement partagées à l'ensemble des plateformes du réseau. Le coût de l'opération (2) est $n * C'_2$, avec n le nombre de plateformes.

Enfin, l'interopération des modèles doit être réalisée. Cependant, la mise en place d'un mécanisme de partage des informations à l'ensemble des nœuds permet une réutilisation des mappings réalisés. Ainsi, chaque modèle ne doit être interopéré qu'une seule fois avec tous les autres modèles, portant le coût de l'opération (3) à $\binom{m}{2} * C'_3$, avec m le nombre de modèles. Le coût total de l'adoption du RDSO est donc $C_{RDSO}(n,m) = n * C'_1 + n * C'_2 + \binom{m}{2} * C'_3 = n * (C'_1 + C'_2) + \binom{m}{2} * C'_3$.

Réduction du coût d'adoption avec le RDSO Supposons que les coûts C_1,C_2,C_3, C'_1, C'_2, C'_3 valent 1 pour permettre une étude comparative des deux formules de coût d'adoption.

Nous avons observé que les coûts des opérations du RDSO sont inférieurs ou égaux aux coûts d'une solution pair à pair. Supposer que tous les coûts sont égaux et valent 1 implique donc une sous-estimation de la réduction du coût d'adoption avec le RDSO.

Il est possible de calculer la formule de réduction du coût d'adoption d'une solution d'échange d'information grâce au RDSO à l'ensemble de la Science Ouverte :

$$R_c(n,m) = \frac{C_{RDSO}(n,m)}{C_{pap}(n,m)} = \frac{2n + \binom{m}{2}}{n + 2 * \binom{n}{2} + n * \binom{m}{2}}$$

La fonction R_c est définie pour $n \in [2; +\infty[$ et $m \in [2; +\infty[$. Observons la limite de cette fonction quand $n \to +\infty$.

$$lim_{n\to+\infty} \frac{2n + \binom{m}{2}}{n + 2 * \binom{n}{2} + n * \binom{m}{2}} = lim_{n\to+\infty} \frac{2n}{2n + 2 * \binom{n}{2}}$$

Nous savons que $2n \leq 2n + \binom{n}{2}$ car $\binom{n}{2} > 0$. De plus, la fonction $\binom{n}{2}$ est croissante monotone. Donc

$$lim_{n\to+\infty} \frac{2n + \binom{m}{2}}{n + 2 * \binom{n}{2} + n * \binom{m}{2}} = 0$$

Observons la limite de cette fonction quand $m \to +\infty$.

$$lim_{m\to+\infty} \frac{2n + \binom{m}{2}}{n + 2 * \binom{n}{2} + n * \binom{m}{2}} = lim_{m\to+\infty} \frac{\binom{m}{2}}{\binom{m}{2}} = 1$$

Nous savons que $n \in [2; +\infty[$

$$lim_{m\to+\infty}\frac{\binom{m}{2}}{n * \binom{m}{2}} \leq lim_{m\to+\infty}\frac{\binom{m}{2}}{2 * \binom{m}{2}}$$

$$lim_{m\to+\infty}\frac{\binom{m}{2}}{n * \binom{m}{2}} \leq \frac{1}{2}$$

Nous observons donc que cette fonction tend vers 0 quand le nombre de plateformes tend vers $+\infty$ et tend vers 0.5 quand le nombre de modèles tend vers $+\infty$. Ainsi, la réduction du coût d'adoption de notre solution tend vers 100% avec un nombre de plateformes croissant et vers 50% avec un nombre de modèles croissant, par rapport à une solution d'échange d'information pair à pair non globalement organisée. Nous observons que $R_c(2, 2) \sim 0.83$. Ainsi, notre solution permet de réduire, quel que soit le nombre de plateformes, le coût d'adoption d'un échange d'information dans la Science Ouverte.

6.2.6 Bilan

Q4 - Est-ce que le RDSO permet de mettre en place un échange d'information interdisciplinaire et intercommunautaire ? Avec la POC du RDSO, les mécanismes décrits dans le chapitre 5 sont implémentables et fonctionnels. Les résultats d'expérimentations nous ont permis d'observer

- Un enrichissement de 7% du volume du jeu de données du Dr. Pressecq et un gain de temps de 80% pour la construction de son jeu de données.
- un échange d'information interdisciplinaire, permettant ainsi le développement de nouvelles questions de recherche mêlant agronomie et sciences sociales pour le Dr. Pressecq.
- un échange d'information intradisciplinaire, permettant de croiser des données provenant des sciences naturelles permettant de penser une approche combinant données des sols de la plateforme Formater et le jeu de données du Dr. Pressecq.

De plus, les données trouvées, qu'elles soient utiles au projet de recherche ciblé ou pour la définition de nouvelles questions de recherche, proviennent de plateformes de communautés différentes, le NCBI faisant partie des plateformes du gouvernement américain et la plateforme data.europa.eu faisant partie des plateformes développées par l'Union Européenne. Ces résultats nous permettent de conclure que **le RDSO permet un échange d'information intradisciplinaire, interdisciplinaire et intercommunautaire.**

Q5 Quelle est l'amélioration du niveau d'échange d'informations dans la Science Ouverte (observé dans la section 3.4.1) apporté par le RDSO ? Nous avons observé dans le cadre de nos expérimentations, qu'avec un faible niveau d'interopération, nous avons réussi à **mettre en place un échange d'information intradisciplinaire, interdisciplinaire et intercommunautaire de 72%**. Ce niveau est atteint en supposant que toutes les plateformes s'intègrent au RDSO et sans amélioration de l'interopération des modèles de métadonnées.

Nous avons estimé que le coût d'adoption du RDSO est bien inférieur au coût d'adoption d'une solution d'échange d'information pair à pair entre les plateformes de la Science Ouverte. Ce coût est réduit de $\sim 17\%$ au minimum et réduit encore avec le nombre de plateformes qui augmente. L'adoption du RDSO à l'ensemble des plateformes de la Science Ouverte est donc une hypothèse probable.

De plus, un enrichissement du registre du RDSO avec plus de modèles de métadonnées et plus de mappings permettrait **une amélioration du niveau d'échange d'information dans la Science Ouverte, théoriquement jusqu'à 100%**. En effet, nous avons validé notre capacité à intégrer les différents modèles de métadonnées de la Science Ouverte, en réalisant une interopération d'une sélection représentative des modèles de métadonnées dans la Science Ouverte.

Ces expérimentations et ces analyses nous permettent de conclure que le RDSO permet de passer le niveau d'échange d'information dans la Science Ouverte de ¡1% à 72%, avec une évolution théorique à 100%. **Le RDSO permet une multiplication au minimum par 100 (de 0.7% à 72%) du niveau d'échange d'information dans la Science Ouverte.**

6.3 Conclusion

Le LDSO, par la réutilisation de l'intégration de métadonnées d'autres plateformes par certains utilisateurs et profitant à l'ensemble des utilisateurs du LDSO permet de réduire le problème de manque d'information. Il s'agit donc d'une solution adaptée à la Science Ouverte à l'échelle communautaire.

L'échange de métadonnées entre les PDRO est nécessaire pour assurer une recherche de données unifiée entre les PDRO. Nous avons proposé deux solutions :

— Le LDSO pour une recherche unifiée intracommunautaire ;
— Le RDSO pour une recherche unifiée interdisciplinaire et intercommunautaire.

Nous avons réalisé une expérimentation pour chacune de ces solutions pour valider leur capacité à atteindre cet objectif et évaluer quantitativement leurs apports. Ces expérimentations ont nécessité de développer des preuves de concepts pour ces deux solutions et de solliciter des utilisateurs.

La première expérimentation concerne le LDSO. Nous avons intégré les données de trois PDRO et nous avons demandé à onze utilisateurs de réaliser des requêtes et de répondre à des questions. Cette expérimentation nous a permis d'observer que :

— La recherche unifiée intracommunautaire de données est bien implantée grâce au LDSO et peut être validée par preuve de concept.
— Le LDSO peut être personnalisable pour s'adapter au plus proche des utilisateurs, notamment en réduisant le nombre d'erreurs de requêtage et en réduisant le temps pour trouver des données de recherche.

La seconde expérimentation porte sur le RDSO. Dans un premier temps, nous avons validé que notre solution permet d'intégrer tous les types de modèles de métadonnées de la Science Ouverte et de s'assurer que notre sélection de données n'introduise pas de biais. Ensuite, nous avons développé une preuve de concept et un protocole d'expérimentation en déployant un RDSO contenant 14 PDRO de toutes les disciplines, gérant une grande variété et un grand volume de données issues de différentes communautés. Nous avons demandé à un chercheur de réaliser les requêtes dans le cadre de son projet de recherche. Les résultats de cette expérimentation nous ont permis d'observer les points suivants :

— Le RDSO permet de réaliser une recherche de données unifiée interdisciplinaire et intercommunautaire, grâce à la preuve de concept et la validation de l'intégration de tous les modèles de métadonnées de la Science Ouverte
— Le RDSO propose une réduction du temps de construction de jeux de données via la recherche de données (80% dans notre exemple) et un enrichissement du volume des données (7% dans notre exemple)

Nous avons proposé deux analyses supplémentaires du RDSO :

— Une évaluation de l'augmentation de l'échange de métadonnées basée sur notre scénario d'expérimentation avec une multiplication par plus de 100 du niveau d'échange de métadonnées, passant de 0.7% à 72% ;

— Une évaluation de la réduction du coût d'adoption par rapport à une solution d'échange de métadonnées pair à pair des PDRO, croissant avec le nombre de plateformes. La valeur minimale de réduction du coût est de 17% du coût d'adoption, dans le cas d'une intégration de deux plateformes et deux modèles de métadonnées différents.

Ces deux solutions se combinent pour offrir une solution complète, offrant un environnement de recherche de données intracommunautaire et de consommation de ces données pour le LDSO et un passage à l'échelle de l'échange de métadonnées permettant une recherche de données intercommunautaire et interdisciplinaire. L'utilisation conjointe de ces solutions permet :

— Au LDSO de combler l'absence d'implantation de mécanisme d'échange de métadonnées passant à l'échelle ;

— Au RDSO d'avoir une architecture pouvant tenir la charge des hubs communautaires.

Ces deux solutions sont complémentaires et proposent une solution sur plusieurs niveaux pour échanger des métadonnées dans la Science Ouverte. Grâce à ces solutions, nous répondons à la problématique d'implantation d'une recherche de données unifiée entre les communautés, les disciplines et les différents acteurs de la Science Ouverte et de réduction du coût de cette recherche de données. Nous avons proposé une solution de réseau de collaboration de chercheurs, qui correspond à la définition de la Science Ouverte.

Les expérimentations sur le LDSO ont fait l'objet d'une publication scientifique dans la conférence internationale ADBIS 2023 (Dang et al. (2023b)). Les expérimentations sur le RDSO ont fait l'objet de deux publications scientifiques dans la conférence internationale RCIS 2024 (Dang et al. (2024a)) et la conférence internationale DASFAA 2024 (Dang et al. (2024b)).

Chapitre 7

Conclusion générale

La **recherche interdisciplinaire** est vue comme un atout permettant de répondre à des questions de recherche trop difficiles à traiter en les approchant par une seule discipline (Corbett et al. (2013)). Cette interdisciplinarité est de plus en plus présente au cœur des questions de recherche abordées par les chercheurs (Ramachandran et al. (2021)). Une partie de l'interdisciplinarité et de son déploiement passe par la collecte de données provenant de différentes disciplines. Cependant, cette collecte de données de recherche nécessite de mettre en place une exploration multi-plateformes et de permettre une recherche de données sur les différentes plateformes existantes, que nous avons appelées Plateformes de Données de Recherche Ouverte (PDRO).

La réponse à cette problématique nécessite la mise en place d'un accès unifié et simplifié à l'ensemble des plateformes et pour cela d'assurer une interopérabilité de celles-ci (Ise (2014)).

7.1 Bilan

7.1.1 Modèle générique de l'interopérabilité

Pour comprendre les enjeux de cette interopérabilité et pouvoir proposer une solution réduisant le coût de recherche de données dans la Science Ouverte, nous avons exploré le concept d'interopérabilité. Nous avons montré avec l'étude de travaux provenant de différentes communautés que l'interopérabilité ne possède aucune compréhension commune (constat partagé par (Rezaei et al. (2014a))). Cette compréhension commune doit être **générique** et **exhaustive** et proposer un référentiel commun de compréhension de l'interopérabilité. Ce référentiel de compréhension est nécessaire pour permettre la spécification de la notion d'interopérabilité à tout sujet de recherche.

Nous avons fait une proposition pour définir ce référentiel générique et exhaustif. L'exhaustivité de notre proposition provient de l'intégration dans notre proposition des concepts suivants :

- une définition générique de l'interopérabilité, des données, de l'information, de l'information utile ainsi que de l'échange d'information et de l'échange de données ;
- une proposition d'un modèle en sept couches pour l'implantation de l'interopérabilité ;
- la définition du triplet de l'interopérabilité permettant l'application de ce modèle d'implantation à un problème spécifique ;
- la définition des différents types de mécanismes d'interopérabilité, avec les mécanismes d'interopérabilité par standardisation et les mécanismes par mise en place de pas-

serelles ;
— des outils permettant l'évaluation quantitative de la qualité de l'implantation de l'interopérabilité.

Nous avons validé la généricité de notre solution en nous assurant de la réponse de notre solution aux critères d'une théorie formelle de l'interopérabilité (Diallo et al. (2011)) ainsi qu'en appliquant notre proposition de compréhension à différents travaux de la littérature.

L'application de notre proposition aux PDRO nous a permis de décomposer l'interopérabilité des PDRO en deux types d'interopérabilité et de simplifier le problème :
— l'interopérabilité des API de communication ;
— l'interopérabilité des systèmes de gestion de métadonnées.

Nous avons ensuite proposé une analyse de l'échange de métadonnées dans la Science Ouverte. Nous avons montré que la probabilité de trouver une donnée recherchée est inférieure à 0.7% avec une confiance de 99,9997%, ce qui montre un **manque d'implantation de l'échange de métadonnées entre les PDRO**. Nous avons aussi montré que les outils de mappings automatiques ne sont pas suffisamment performants pour résoudre la problématique de variété des modèles de métadonnées. Ainsi, pour permettre la mise en place d'un réseau de collaboration entre chercheurs (cf définition de la Science Ouverte par Vicente-Saez and Martinez-Fuentes (2018)), nous devrons faciliter une recherche unifiée et proposer par une réponse architecturale (cf les étapes pour la mise en place de la Science Ouverte par Nosek (2019)).

7.1.2 Recherche et consommation de données intracommunautaires

Nous avons proposé une extension du concept de lac de données avec le Lac de Données de la Science Ouverte (LDSO) via la définition de ce nouveau concept et l'explicitation des architectures fonctionnelle et technique. Le LDSO se base sur :
— une gestion multi-modèles de métadonnées, basée sur des mappings entre les modèles afin de fournir un accès unifié aux métadonnées ;
— un ajout de mécanismes de contrôle d'accès et d'authentification permettant aux propriétaires de données un contrôle sur l'ouverture de celles-ci ;
— une intégration virtuelle de données permettant de pallier le trop grand volume dans la Science Ouverte.

Pour l'implantation de l'interopérabilité des PDRO, le LDSO se base sur :
— Une approche d'interopérabilité hybride pour répondre à la variété des API, avec un mécanisme d'interopérabilité par standardisation avec une grande partie de PDRO en choisissant une API REST et un mécanisme manuel d'interopérabilité par mise en place de passerelles avec les autres types d'API, permis par les API REST.
— Une interopérabilité par mise en place de passerelles pour répondre à la variété des modèles de métadonnées, avec la mise en place de mappings manuels entre les concepts des modèles de métadonnées.

Cette solution permet de proposer une recherche locale unifiée intracommunautaire de données de recherche. Elle offre aussi un espace d'analyse pour les chercheurs, leur permettant la réalisation d'apprentissage de modèles d'intelligence artificielle, de jointure de jeux de données pour des analyses croisées ou la définition de revues systématiques.

7.1.3 Exploration de données de recherche interdisciplinaire et intercommunautaire

Les contraintes apportées par la Science Ouverte, en termes de volumétrie (de plateformes, des données) et de variété (de modélisation de métadonnées), ne permettent pas à des solutions centralisées de voir le jour pour répondre à la problématique. Pour assurer une recherche interdisciplinaire et intercommunautaire de données de recherche, nous avons proposé une solution de réseau d'interconnexion de plateformes décentralisé, distribué et fédéré avec le Réseau de Données de la Science Ouverte (RDSO). Ce réseau est conçu pour proposer une recherche unifiée de données et pour être résilient aux pannes et aux cyberattaques afin d'assurer une sécurisation de cette recherche de données.

Pour implanter l'interopérabilité des PDRO, le RDSO se base sur :

— Une interopération par standardisation pour les API de communication et pouvant être enrichie si besoin par des mécanismes d'interopérabilité par mise en place de passerelles grâce aux API REST ;

— Une interopération par mise en place de passerelles pour les modèles de métadonnées, grâce à l'utilisation de mappings dans le registre.

7.1.4 Evaluation de nos solutions

Nous avons proposé une expérimentation pour valider chaque solution. Chaque expérimentation déploie une preuve de concept qui sert de support à une expérience utilisateur. L'ensemble du code est disponible dans un dépôt Github public (cf Chapitre 6). Nous avons observé que :

— le LDSO permet un gain de temps proportionnel au nombre de PDRO externes gérées par le LDSO ;

— le LDSO fournit un outil personnalisable adapté au plus près des besoins des chercheurs et une solution facile à prendre en main pour les chercheurs.

L'expérimentation du RDSO est composée du développement d'une preuve de concept accompagnée d'une vidéo de démonstration suivie d'une expérience utilisateur avec un projet de recherche réel. Les avantages de notre solution sont les suivants :

— le RDSO a permis un échange interdisciplinaire et intercommunautaire ;

— le RDSO a permis une réduction du temps nécessaire à la construction d'un jeu de données de 80%, un enrichissement du volume de données de 7% ;

— le RDSO a permis au Dr. Pressecq de concevoir de nouvelles questions intradisciplinaires et interdisciplinaires grâce aux jeux de données trouvés sur le RDSO dont il n'avait pas connaissance ;

— le RDSO propose une amélioration de l'échange de métadonnées de 0.7% à 72%, représentant une multiplication par plus de 100 de ce niveau d'échange de métadonnées ;

— le RDSO propose une grande réduction du coût d'adoption par rapport à une solution pair à pair, croissant avec le nombre de plateformes et représentant une réduction du coût d'adoption minimum de 17%.

Ces expérimentations montrent que nous avons proposé une solution complète de recherche de données unifiée interdisciplinaire, intercommunautaire et intracommunautaire. Le RDSO et le LDSO sont deux solutions complémentaires permettant la proposition d'une réponse à deux échelles.

7.2 Perspectives de recherche

Ces solutions répondent à la problématique de la mise en place d'un réseau de collaboration basé sur une recherche unifiée de données. Plusieurs pistes sont prévues pour les améliorer.

7.2.1 Perspectives à court terme

Notre proposition s'est concentrée sur la recherche de données. Le RDSO propose une recherche de données unifiée à l'ensemble des PDRO inscrites dans le RDSO. Cependant, cette recherche ne permet pas le téléchargement des données. Une extension du processus de requêtage du RDSO permettrait d'intégrer des requêtes auprès des systèmes de gestion de données des PDRO pour la récupération des données.

Les approches "polystores" (Bondiombouy and Valduriez (2016)) sont les plus propices à l'enrichissement du RDSO au vu de l'hétérogénéité des sources et l'approche distribuée dans le RDSO. L'approche type "polystore" permettrait d'enrichir le système de requêtage du RDSO pour intégrer la récupération automatique des données. En s'appuyant sur le langage de requêtage proposé par El Ahdab et al. (2023a), nous envisageons d'étendre l'interrogation pour intégrer la récupération des données par une interrogation des systèmes de gestion de données des PDRO. Ce langage de requêtage doit être intégré dans le processus de propagation des requêtes. La vue unifiée proposée par El Ahdab et al. (2023b) correspond dans le RDSO au registre distribué. Le registre distribué doit être enrichi pour permettre la conservation des informations nécessaires à cette extension, notamment pour les opérateurs du langage de requêtage proposé.

Nous avons observé dans le chapitre 5 que le RDSO diffère en quatre points de la maille de données (l'objectif, la portée de l'interopérabilité, le quantum architectural et la gouvernance). Pour répondre au même objectif que la maille de données, il convient d'ajouter au RDSO la capacité de récupérer des données et de permettre une consommation de données dans le RDSO.

7.2.2 Perspectives à moyen terme

Les mappings utilisés dans les deux solutions définissent une relation d'égalité entre deux concepts. Cependant, les mappings à gérer sont plus riches et nécessitent de pouvoir exprimer des relations plus complexes qu'une relation binaire d'égalité entre deux concepts.

La campagne d'évaluation d'OAEI [1] montre que les mappings entre modèles peuvent être complexes. Ces mappings complexes intègrent des constructeurs ou des opérateurs de transformation (Thiéblin et al. (2020)). La prise en compte de plusieurs travaux est envisagée :

— les travaux sur les mappings n-aires pour étendre l'arité des mappings dans le RDSO (Moran et al. (2009)) ;
— les travaux sur le mapping holistique (Megdiche (2015)) ou de "network matching" (Santos and Mello (2022)) pour améliorer les performances de mappings automatiques et réduire encore le coût d'adoption du RDSO.

1. `https://oaei.ontologymatching.org/2023/complex/index.html`

L'implantation de la richesse des relations sémantiques du Semantic Mapping Vocabulary [2] permettrait de capturer cette finesse des relations entre les informations. Cet enrichissement se base sur une extension du format EDOAL [3] afin d'inclure cette diversité dans le RDSO.

L'ensemble de ces aspects seront intégrés dans le RDSO et le LDSO pour proposer une recherche de données plus riche.

7.2.3 Perspectives à long terme

L'objectif du RDSO est la recherche transparente et interdisciplinaire de données. Pour améliorer cette solution, nous prévoyons un enrichissement de l'objectif avec une consommation de données transparente. L'adaptation de l'approche de federated learning (El Rifai et al. (2020)) offre une approche de consommation distribuée au sein du RDSO. Elle permettrait de profiter de l'ensemble des ressources existantes dans la Science Ouverte. Cette approche pourrait proposer des processus de consommation sur les données là où elles se trouvent, réduisant la quantité de données échangées mais aussi de profiter de la puissance de calcul existante dans la Science Ouverte.

Cette approche s'adapte naturellement à la mise en place d'apprentissage de modèles d'intelligence artificielle El Rifai et al. (2020). Dans le but de réaliser des revues systématiques, plusieurs processus peuvent être exécutés sur des plateformes proposant des publications scientifiques :

— Une détection des publications fausses (Labbé and Labbé (2013)) ou contenant des expressions torturées (Cabanac et al. (2021)) afin d'augmenter la qualité des données trouvées ;
— La synthèse automatique de documents (El-Kassas et al. (2021)) afin de réduire le coût de la sélection de données.

Pour faciliter la jointure de données, l'intégration d'algorithmes de data mining distribués (Zeng et al. (2012)) permettrait de fournir un croisement de jeux de données provenant de l'ensemble des plateformes du RDSO.

2. https://mapping-commons.github.io/semantic-mapping-vocabulary/
3. https://moex.gitlabpages.inria.fr/alignapi/edoal.html

Bibliographie

Ambrosio, R. and Widergren, S. (2007). A framework for addressing interoperability issues. In *2007 IEEE Power Engineering Society General Meeting*, pages 1–5. IEEE.

Aryani, A., Fenner, M., Manghi, P., Mannocci, A., and Stocker, M. (2020). Open science graphs must interoperate ! In *International Conference on Theory and Practice of Digital Libraries*, pages 195–206. Springer.

Aydinoglu, A. U., Suomela, T., and Malone, J. (2014). Data management in astrobiology : Challenges and opportunities for an interdisciplinary community. *Astrobiology*, 14(6) :451–461.

Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge University Press, Cambridge.

Beheshti, A., Benatallah, B., Nouri, R., and Tabebordbar, A. (2018). Corekg : a knowledge lake service. *Proceedings of the VLDB Endowment*, 11(12) :1942–1945.

Berre, A. J., Elvesæter, B., Figay, N., Guglielmina, C., Johnsen, S. G., Karlsen, D., Knothe, T., and Lippe, S. (2007). The athena interoperability framework. In *Enterprise interoperability II : new challenges and approaches*, pages 569–580. Springer.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008.

Bondiombouy, C. and Valduriez, P. (2016). Query processing in multistore systems : an overview. *International Journal of Cloud Computing*, 5(4) :309–346.

Cabanac, G., Labbé, C., and Magazinov, A. (2021). Tortured phrases : A dubious writing style emerging in science. evidence of critical issues affecting established journals. *arXiv preprint arXiv :2107.06751*.

Castro, J. P. d. C., Romero, L. d. M. F., Carniel, A. C., and Aguiar, C. D. d. (2022). Fair principles and big data : A software reference architecture for open science. In *Proceedings*.

Corbett, C. F., Costa, L. L., Balas, M. C., Burke, W. J., Feroli, E. R., and Daratha, K. B. (2013). Facilitators and challenges to conducting interdisciplinary research. *Medical care*, 51 :S23–S31.

Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M., van de Sanden, M., and Coppens, F. (2021). Eosc interoperability framework. *Report from the EOSC Executive Board Working Groups FAIR and Architecture*, 10 :620649.

Daemen, J. and Rijmen, V. (1999). Aes proposal : Rijndael.

Dang, V.-N., Aussenac-Gilles, N., Megdiche, I., and Ravat, F. (2023a). Interoperability of open science metadata : What about the reality ? In *International Conference on Research Challenges in Information Science*, pages 467–482. Springer.

Dang, V.-N., Aussenac-Gilles, N., Megdiche, I., and Ravat, F. (2024a). Enabling interdisciplinary research in open science : Open science data network. In *International Conference on Research Challenges in Information Science*, pages 19–34. Springer.

Dang, V.-N., Aussenac-Gilles, N., Megdiche, I., and Ravat, F. (2024b). Osdn : an open science data network for interdisciplinary research. In *29th International Conference on Database Systems for Advanced Applications, DASFAA 2024*, volume 14856. Springer Singapore.

Dang, V.-N., Aussenac-Gilles, N., and Ravat, F. (2023b). Multi-disciplinary research : Open science data lake. In *European Conference on Advances in Databases and Information Systems*, pages 71–81. Springer.

Dehghani, Z. (2019). How to move beyond a monolithic data lake to a distributed data mesh. *Martin Fowler's Blog*, page 45.

Dehghani, Z. (2020). Data mesh principles and logical architecture. *martinfowler. com*.

Diallo, S. Y. et al. (2011). Understanding interoperability. In *Proceedings of the 2011 Emerging M&S Applications in Industry and Academia Symposium*, pages 84–91.

Dolhopolov, A., Castelltort, A., and Laurent, A. (2023a). Exploring the benefits of blockchain-powered metadata catalogs in data mesh architecture. In *International Conference on Management of Digital*, pages 32–40. Springer.

Dolhopolov, A., Castelltort, A., and Laurent, A. (2023b). Implementing a blockchain-powered metadata catalog in data mesh architecture. In *International Congress on Blockchain and Applications*, pages 348–360. Springer.

Dolhopolov, A., Castelltort, A., and Laurent, A. (2023c). Trick or treat : Centralized data lake vs decentralized data mesh. In *International Conference on Management of Digital*, pages 303–316. Springer.

Dooley, R., Brandt, S. R., and Fonner, J. (2018). The agave platform : An open, science-as-a-service platform for digital science. In *Proceedings of the Practice and Experience on Advanced Research Computing*, pages 1–8.

Dutta, B. and Patel, J. (2021). Amv : Algorithm metadata vocabulary. *arXiv preprint arXiv :2106.03567*.

Dymytrova, V. and Paquienséguy, F. (2017). Analyse de portails métropolitains de données ouvertes à l'échelle internationale. *Livrable 2*.

El Ahdab, L., Teste, O., Megdiche, I., and Péninou, A. (2023a). A polystore querying system applied to heterogeneous and horizontally distributed data. In *International Conference on Database and Expert Systems Applications*, pages 437–442. Springer.

El Ahdab, L., Teste, O., Megdiche, I., and Péninou, A. (2023b). Unified views for querying heterogeneous multi-model polystores. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 319–324. Springer.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization : A comprehensive survey. *Expert systems with applications*, 165 :113679.

El Rifai, O., Biotteau, M., de Boissezon, X., Megdiche, I., Ravat, F., and Teste, O. (2020). Blockchain-based federated learning in medicine. In *Artificial Intelligence in Medicine : 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pages 214–224. Springer.

Ermilov, I., Höffner, K., Lehmann, J., and Mouromtsev, D. (2015). kore : Using linked data for openscience information integration. In *SEMANTiCS (Posters & Demos)*, pages 67–70.

European Commission and Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M., Choirat, C., Sanden, M., and Coppens, F. (2021). *EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture*. Publications Office.

Gleizes, M.-P., Boes, J., Lartigue, B., and Thiébolt, F. (2018). neocampus : A demonstrator of connected, innovative, intelligent and sustainable campus. In *Intelligent Interactive Multimedia Systems and Services 2017 10*, pages 482–491. Springer.

Gonçalves, R. S. et al. (2019). Aligning biomedical metadata with ontologies using clustering and embeddings. In *European Semantic Web Conference*, pages 146–161. Springer.

Ise, O. A. (2014). Towards a unified university information system : bridging the gap of data interoperability. *American Journal of Software Engineering*, 2(2) :26–32.

Kathawalla, U.-K., Silverstein, P., and Syed, M. (2021). Easing into open science : A guide for graduate students and their advisors. *Collabra : Psychology*, 7(1) :18684.

Killough, B. (2018). Overview of the open data cube initiative. In *IGARSS 2018-2018 IEEE international geoscience and remote sensing symposium*, pages 8629–8632. IEEE.

Kostoska, M., Gusev, M., and Ristov, S. (2016). An overview of cloud interoperability. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 873–876. IEEE.

Koussouris, S., Lampathaki, F., Mouzakitis, S., Charalabidis, Y., and Psarras, J. (2011). Digging into the real-life enterprise interoperability areas definition and overview of the main research areas. *Proceedings of CENT*, pages 19–22.

Koutras, C. et al. (2021). Valentine : Evaluating matching techniques for dataset discovery. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 468–479. IEEE.

Labbé, C. and Labbé, D. (2013). Duplicate and fake publications in the scientific literature : how many scigen papers in computer science ? *Scientometrics*, 94 :379–396.

Li, Z. et al. (2021). Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 408–417.

Maciel, R. S., Valle, P. H., Santos, K. S., and Nakagawa, E. Y. (2024). Systems interoperability types : A tertiary study. *ACM Computing Surveys*.

Massmann, S., Engmann, D., and Rahm, E. (2006). Coma++ : Results for the ontology alignment contest oaei 2006. *Ontology Matching*, 225.

Megdiche, I. (2015). *Intégration holistique et entreposage automatique des données ouvertes.(Holistic integration and automatic warehousing of open data).* PhD thesis, Paul Sabatier University, Toulouse, France.

Moran, K., Claypool, K. T., and Hescott, B. J. (2009). Compositematch : Detecting n-ary matches in ontology alignment. In *OM*.

Nilsson, M., Baker, T., and Johnston, P. (2008). Interoperability levels for dublin core metadata. Technical report, Dublin Core Metadata Initiative.

Nosek, B. (2019). Strategy for culture change. *Center for Open Science*, 11.

Noura, M. et al. (2019). Interoperability in internet of things : Taxonomies and open challenges. *Mobile networks and applications*, 24(3) :796–809.

Oh, H., Jones, A., Finin, T., et al. (2024). Employing word-embedding for schema matching in standard lifecycle management. *Journal of Industrial Information Integration*, 38 :100547.

Peisert, S., Welch, V., Adams, A., Bevier, R., Dopheide, M., LeDuc, R., Meunier, P., Schwab, S., and Stocks, K. (2017). Open science cyber risk profile (oscrp), version 1.3.3.

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., and Peres-Neto, P. (2019). Ecological data should not be so hard to find and reuse. *Trends in ecology & evolution*, 34(6) :494–496.

Pressecq, T. (2023). *Développement d'outil d'aide à la décision pour favoriser l'usage du biocontrôle microbien.* Theses, Université d'Avignon.

Rainey, L., Lutomski, J. E., and Broeders, M. J. (2023). Fair data sharing : An international perspective on why medical researchers are lagging behind. *Big Data & Society*, 10(1) :20539517231171052.

Ramachandran, R., Bugbee, K., and Murphy, K. (2021). From open data to open science. *Earth and Space Science*, 8(5) :e2020EA001562.

Ravat, F. and Zhao, Y. (2019a). 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*, pages 304–313. Springer.

Ravat, F. and Zhao, Y. (2019b). Metadata management for data lakes. In *New Trends in Databases and Information Systems : ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23*, pages 37–44. Springer.

Rezaei, R., Chiew, T. K., and Lee, S. P. (2014a). A review on e-business interoperability frameworks. *Journal of Systems and Software*, 93 :199–216.

Rezaei, R., Chiew, T. K., Lee, S. P., and Shams Aliee, Z. (2014b). Interoperability evaluation models : A systematic review. *Computers in Industry*, 65(1) :1–23.

Sadeh, Y., Denejkina, A., Karyotaki, E., Lenferink, L. I., and Kassam-Adams, N. (2023). Opportunities for improving data sharing and fair data practices to advance global mental health. *Cambridge Prisms : Global Mental Health*, 10 :e14.

Santos, F. and Mello, C. E. (2022). Matching network of ontologies : a random walk and frequent itemsets approach. *IEEE Access*, 10 :44638–44659.

Tanhua, T., Pouliquen, S., Hausman, J., O'brien, K., Bricher, P., De Bruin, T., Buck, J. J., Burger, E. F., Carval, T., Casey, K. S., et al. (2019). Ocean fair data services. *Frontiers in Marine Science*, 6 :440.

Thiéblin, E., Haemmerlé, O., Hernandez, N., and Trojahn, C. (2020). Survey on complex ontology matching. *Semantic Web*, 11(4) :689–727.

Tolk, A., Diallo, S. Y., and Turnitsa, C. D. (2007). 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).

Top, J., Janssen, S., Boogaard, H., Knapen, R., and Şimşek-Şenel, G. (2022). Cultivating fair principles for agri-food data. *Computers and Electronics in Agriculture*, 196 :106909.

Ulrich, H. et al. (2022). Understanding the nature of metadata : systematic review. *Journal of medical Internet research*, 24(1) :e25440.

Uzwyshyn, R. (2016). Research data repositories : the what, when, why and how.

Van Der Veer, H. and Wiles, A. (2008). Achieving technical interoperability. *European telecommunications standards institute*.

Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., and Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC public health*, 14 :1–9.

Van Steen, M. and Tanenbaum, A. S. (2023). *Distributed systems*. Maarten van Steen Leiden, The Netherlands, 4ème edition.

Vicente-Saez, R. and Martinez-Fuentes, C. (2018). Open science now : A systematic literature review for an integrated definition. *Journal of business research*, 88 :428–436.

Vuong, Q.-H., La, V.-P., Vuong, T.-T., Ho, M.-T., Nguyen, H.-K. T., Nguyen, V.-H., Pham, H.-H., and Ho, M.-T. (2018). An open database of productivity in vietnam's social sciences and humanities for public use. *Scientific data*, 5(1) :1–15.

Wang, W., Tolk, A., and Wang, W. (2009). The levels of conceptual interoperability model : applying systems engineering principles to m&s. *arXiv preprint arXiv :0908.0191*.

Wider, A., Verma, S., and Akhtar, A. (2023). Decentralized data governance as part of a data mesh platform : concepts and approaches. In *2023 IEEE International Conference on Web Services (ICWS)*, pages 746–754. IEEE.

Wilkinson, M. D. et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1) :1–9.

Zeng, L., Li, L., Duan, L., Lu, K., Shi, Z., Wang, M., Wu, W., and Luo, P. (2012). Distributed data mining : a survey. *Information Technology and Management*, 13 :403–409.

Zeng, M. L. (2019). Interoperability. *KO Knowledge Organization*, 46(2) :122–146.

Zhao, Y. (2021). *Metadata management for data lake governance*. PhD thesis, Toulouse 1.

Zhao, Y., Megdiche, I., and Ravat, F. (2021). Data lake ingestion management. *arXiv preprint arXiv :2107.02885*.

Zwegers, A. (2003). Ideas roadmap for ebusiness interoperability. In *eGovernment Interoperability Workshop, Brussels*.