



République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderrahmane Mira de Bejaia
Faculté des Sciences Exactes
Département d'Informatique

ECOLE DOCTORALE RESEAUX ET SYSTEMES DISTRIBUES

Mémoire de Magistère

En Informatique

Option : Réseaux et Systèmes Distribués

Thème

Les services de la bibliothèque numérique « Digital Library » et leur interopérabilité basés sur la technologie GRID

Présenté par

Amel BOUFENISSA

Devant le jury composé de :

Président	Abdelkamel TARI	MC/A	Université de Bejaïa
Rapporteur	Nadjib BADACHE	Professeur	CERIST
Examineur	Omar NOUALI	DR	CERIST
Examineur	Makhlouf ALIOUAT	MC/A	Université de Sétif
Invité	Aouaouache EL MAOUHAB	CR	CERIST

Promotion : 2009/2010

REMERCIEMENTS

C'est avec un réel plaisir et un grand enthousiasme que je me livre à la rédaction de cette page. Bien plus que le point final du manuscrit, cette page constitue l'opportunité de m'accorder une réflexion sur une période de ma vie très riche en émotions.

Je remercie en premier notre grand Dieu pour m'avoir donné le courage et la volonté pour terminer ce modeste travail.

Un grand merci à mon directeur de thèse, le Professeur Nadjib BADACHE, directeur du CERIST. Je lui suis reconnaissante de m'avoir donné l'opportunité de préparer mon diplôme de magistère.

Je tiens à exprimer ma très vive reconnaissance envers Madame Aouaouache EL-MAOUIHAB, chargée de recherche au CERIST, pour la qualité de son encadrement. Ses compétences scientifiques, mais aussi ses qualités humaines ont été des éléments précieux pour l'avancement de mon travail. Je lui suis reconnaissante pour sa patience, ses encouragements, sa sympathie et sa disponibilité à tout moment.

Je suis vivement reconnaissante à Monsieur Abdelkamel TARI d'avoir accepté la charge de président du jury.

Je tiens à exprimer mes sincères remerciements aux Messieurs Omar NOUALI et Makhoulf ALIOUAT, qui m'ont fait l'honneur d'évaluer ce travail et qui, par la qualité et la pertinence de leurs remarques, permettront d'améliorer la rédaction de cette thèse. Je les remercie sincèrement de m'avoir permis de profiter de leurs compétences et de leurs connaissances, et l'honneur qu'ils m'ont accordé en acceptant la charge de rapporter mon travail.

J'exprime également toute ma gratitude à ma très chère amie Faiza, pour son aide précieuse, ses encouragements et sa disponibilité durant toute la durée de préparation de mon mémoire.

Je remercie très particulièrement ma chère maman et mon mari pour m'avoir soutenu et encouragé pendant toute la durée de la préparation de cette thèse. Leur soutien et amour sans limites ont été essentiels. Cette thèse est aussi le fruit de leur effort.

Un grand MERCI à Mon Papa, Ma sœur et Mes frères, Mes tantes et Mes oncles pour leurs encouragements afin que ce travail arrive à sa fin.

Je remercie aussi mes beaux-parents pour m'avoir encouragé.

Je ne pourrais terminer ces remerciements sans y associer mes amis du CERIST, en particulier, Lamia, Houda et Sahar, et tant d'autres sans leur soutien je n'aurais pu entreprendre ces études.

Et pour être sûr de n'oublier personne, que tous ceux, qui de près ou de loin, ont contribué par leurs conseils, leurs encouragements ou leur amitié, à l'aboutissement de ce modeste travail, trouvent ici l'expression de ma profonde reconnaissance.

RÉSUMÉ

Plusieurs bibliothèques numériques utilisent le modèle de la recherche distribuée. Sur un certain nombre de sites, les modèles restent satisfaisants. A partir d'un nombre plus important, le modèle montre ses limites et d'autres modèles s'y prêtent mieux. Le plus adapté reste le modèle « harvest metadata » qui est basé sur l'Open Archive Infrastructure / PMH. Dans un contexte d'une collection de bibliothèques numériques, les performances d'une telle infrastructure qui repose sur la découverte de ressources informationnelles et sur l'exécution d'indexation nécessitent de considérer une base architecturale basée sur la haute performance.

La haute performance est une contrainte qui devient de plus en plus solvable sur des architectures de grilles de calcul et de stockage. Dans des environnements dynamiques et hétérogènes, les architectures GRID offre des fonctionnalités de partage de ressources, de stockage et de gestion des données distribuées sur ces clusters GRID.

A cet effet, l'objet de ce travail est de montrer la faisabilité de l'intégration de la technologie des OAI avec leur interopérabilité dans un contexte large avec la technologie GRID. Pour cela, nous avons donc étudié les technologies associées à cette thématique et validé l'approche d'intégration par une implémentation sur un environnement GRID.

MOTS-CLÉS : Bibliothèques numériques, archives ouvertes, interopérabilité, OAI-PMH, grille de calcul.

ABSTRACT

Several digital libraries use the model of distributed research. On a certain number of sites, the model still satisfied. From a larger number, the model shows its limits and other models are better suited. The most suitable model is the "harvest metadata" that is based on the Open Archive Infrastructure / PMH. In the context of a digital library collection, the performance of such infrastructure based on the discovery of informational resources and the execution of indexation requires considering a basic architecture based on high performance.

High performance is a constraint becoming more creditworthy in grid computing and storage architectures. In dynamic environments and heterogeneous, GRID architectures provides functionality for resource sharing, storage and management of data distributed over these GRID clusters.

For this purpose, the purpose of this work is to demonstrate the feasibility of integrating OAI technology with their interoperability in a broad context with the GRID technology. For this, we have studied the technologies associated with this theme and validated the approach of integration with an implementation on a GRID environment.

KEYWORDS: Digital libraries, open archives, interoperability, OAI-PMH, grid computing.

INTRODUCTION GÉNÉRALE	1
1. Introduction.....	1
2. Problématique	1
3. Contribution	2
4. Organisation du document	3
CHAPITRE 1 : LES ARCHIVES OUVERTES	4
1. Introduction.....	4
2. Le système de la communication scientifique	4
3. Le mouvement pour le libre accès à la recherche	4
4. Les archives ouvertes.....	6
4.1. L'Objectif des archives ouvertes.....	7
4.2. Le contenu d'une archive ouverte	7
4.3. Les différents types d'archives ouvertes	8
4.4. Les acteurs d'une archive ouverte.....	8
4.4.1. L'administrateur de l'archive ou le modérateur	8
4.4.2. L'auteur	8
4.4.3. Le lecteur	9
4.4.4. Liens entre les acteurs	9
5. Aspect juridique	9
5.1. Les archives ouvertes et le droit.....	10
6. Les bibliothèques et les archives ouvertes	10
6.1. Un dispositif d'accompagnement dans la transition vers le libre accès.....	10
6.2. Les professionnels de la documentation et les archives ouvertes	11
7. Interopérabilités des données	11
7.1. Z39.50	12
7.2. SRU	12
7.3. OAI-PMH.....	13
7.4. Recherche croisée VS collecte de métadonnées	13
8. Quelques plates-formes de création/gestion d'archive ouverte	14
8.1. Archimede	14
8.2. DSpace	14
8.3. Eprints	14
9. Conclusion	15
CHAPITRE 2 : LE PROTOCOLE OAI-PMH	16
1. Introduction.....	16
2. Assurer l'interopérabilité des données.....	16

2.1. Le concept de métadonnées	17
2.2. Le Dublin Core (DC)	18
3. Le protocole OAI-PMH	20
3.1. Les acteurs.....	20
3.1.1. Les fournisseurs de données (DP : data providers)	20
3.1.2. Les fournisseurs de service (SP : Service providers)	21
3.1.3. L'agrégateur, fournisseur de données intermédiaire	21
3.2. Principes de fonctionnement du protocole OAI.....	22
3.2.1. Les concepts	22
3.2.2. Principes organisationnels	23
3.2.3. L'architecture technologique du protocole OAI	24
3.3. Concevoir des services OAI.....	25
3.3.1. L'entrepôt OAI	26
3.3.2. Le moissonneur OAI	27
4. Exemples d'entrepôts et moissonneurs OAI.....	27
4.1. Entrepôts OAI	27
4.2. Moissonneurs OAI	28
5. Conclusion	28
CHAPITRE 3 : LES GRILLES DE CALCUL	29
1. Introduction.....	29
2. Origine	29
3. Définitions	30
4. Principes communs des grilles.....	31
5. Classement et exemples de grilles	31
5.1. Les grilles d'information	32
5.2. Les grilles de stockage	32
5.3. Les grilles de calcul.....	32
6. Les applications concernées par les grilles informatiques	33
7. Architecture de la grille	34
7.1. Architecture Selon Foster et al.....	34
7.1.1. La couche applicative	34
7.1.2. La couche collective	34
7.1.3. La couche ressources.....	35
7.1.4. La couche connexion.....	35
7.1.5. La couche fabrique	35
8. Intergiciel	36
8.1. Définition	36
8.2. L'architecture d'un intergiciel.....	36
9. Organisations virtuelles	37
10. Fonctionnement d'une grille	38

10.1. Composants gLite	39
10.2. Fonctionnement interne de gLite	39
10.2.1. Service d'information (SI)	39
10.2.2. Le service de gestion de la charge de travail (WMS).....	40
10.2.3. Le langage de description de job (JDL)	41
10.2.4. L'élément de calcul (CE)	45
10.2.5. La gestion des données (DM).....	46
10.3. Logging and BookKeeping (LB)	48
10.4. Mécanisme de sécurité dans gLite	49
10.5. Utilisation de gLite :.....	50
10.5.1. Chemins d'un job : de la soumission à la collection.....	50
10.5.2. Les différents états d'un job soumis à la grille.....	51
11. OAI et Grille de calcul.....	52
12. Conclusion	53
CHAPITRE 4 : APPROCHE PROPOSÉE	55
1. Introduction.....	55
2. Le moissonnage des métadonnées	55
2.1. Cadre conceptuel de l'approche proposée.....	56
2.1.1. Les acteurs de l'approche proposée.....	56
2.1.2. Les étapes du processus de collecte.....	57
2.2. Architecture du système proposé	59
2.2.1. Couche fournisseur de données	59
2.2.2. Couche collecte et filtrage	60
2.2.2.1. Service de collecte d'URLs OAI	61
2.2.2.2. Service de filtrage de métadonnées	63
2.2.3. Couche moissonneur	65
2.2.3.1. Nœuds de moissonnage	66
2.2.3.2. Service de planification du moissonnage	67
2.2.4. Couche référentiel de métadonnées.....	73
3. La recherche.....	73
4. Conclusion	74
CHAPITRE 5 : VALIDATION ET EXPÉRIMENTATION	42
1. Introduction.....	72
2. Environnement de développement.....	72
2.1. Le middleware g-Lite	72
2.2. Open Harvester Systems	72
2.2.1. Fonctionnalités de OHS.....	73
2.2.2. Vue d'ensemble du système OHS	73
2.2.3. Composants du système OHS	74
2.2.4. Schémas de la base de données d'OHS	75
3. Implémentation du modèle proposé.....	77
3.1. Modification de l'outil OHS	77

3.2. Langages de développement	78
3.2.1. JAVA.....	78
3.2.2. JDL	79
3.3. Contraintes et solutions	79
3.4. Utilisation de gLite.....	80
3.4.1. Initialisation.....	80
3.4.2. Soumission de job (Job Submission).....	80
3.4.3. État du job (Job Status)	82
3.4.4. Annulation d'un job	83
3.4.5. Collecte de résultats pour un job	83
4. Expérimentation.....	84
6. Discussion.....	88
5. Conclusion	89
CONCLUSION GÉNÉRALE	89