Action Concertée Incitative
[ACI]
Globalisation des Ressources
Informatiques et des Données
[GRID]

Rapport de fin de projet retenu dans le cadre de la campagne ACI GRID 2002

RAPPORT DE FIN DE PROJET

.....

PADOUE (Partage de Données Utiles en Environnement)

.....

Par

.....Anne Doucet.....

Décembre 2005

Adresse du Laboratoire porteur du projet :

.....

LIP6

8 rue du Capitaine Scott

75015 Paris

Action Concertée Incitative
[ACI]
Globalisation des Ressources
Informatiques et des Données
[GRID]

Rapport de fin de projet retenu dans le cadre de la campagne ACI GRID 2002

1 PARTICIPANTS

- Laboratoire d'Informatique de Paris 6 (LIP6)
  - Porteur du projet (tâche 6)
  - Responsable de l'exploitation des méta-données (tâche 3)

- Institut National de Recherche en Informatique et en Automatique (INRIA)
  - Responsable de la gestion des chaînes de traitement (tâche 1)
  - Responsable de la médiation d'information (tâche 4)

- Laboratoire d'Ingénierie pour les Systèmes Complexes (LISC- Cemagref)
  - Responsable de la validation expérimentale (tâche 5)

- Laboratoire d'Informatique de Robotique, de Microélectronique de Montpellier (LIRMM)
  - Responsable (avec l'IRD) de la gestion des méta-données (tâche 2)

- Structures et Systèmes Spatiaux (UMR 3S, Cemagref, Montpellier)
  - Responsable (avec l'IRD) des spécifications et de l'architecture (tâche 0)

- Institut de Recherche pour le Développement (IRD Montpellier)
  - Responsable (avec l'UMR 3S) des spécifications et de l'architecture (tâche 0)
  - Responsable (avec le LIRMM) de la gestion des méta-données (tâche 2)

- Centre de Données astronomiques de Strasbourg (CDS)

2 OBJECTIFS DU PROJET

Les chercheurs scientifiques et en particulier les chercheurs de l'environnement, disposent de grandes quantités de données stockées dans des sources d'information autonomes, fortement hétérogènes et dispersées géographiquement, qui potentiellement représentent un formidable réseau de ressources partageables. Parallèlement, la communauté scientifique a développé une multitude de programmes informatiques d'analyse et de traitement des données jouant un rôle essentiel dans les activités de surveillance de l'environnement ou dans les processus décisionnels qui doivent se baser sur des prédictions de qualité. Le problème majeur auquel sont confrontés les organismes de recherche est l'exploitation de cet ensemble de ressources

L'objectif du projet PADOUE est de développer une architecture de partage de données environnementales, qui doit permettre aux chercheurs de l'environnement de mutualiser leurs ressources et de les exploiter, afin de mettre à la disposition des personnes en charge de la surveillance ou de la prise de décision l'information de la meilleure qualité possible.

L'exploitation de ces ressources nécessite de savoir quelles sont les ressources existantes, où elles se trouvent, comment les rendre interopérables, et comment les utiliser (avec quels programmes, quels jeux de données).

Le projet PADOUE mène donc quatre actions complémentaires de recherche visant à :

- assurer l'interopérabilité des données et des programmes de traitement. Cette action s'appuiera sur le système de médiation LeSelect (développé dans le projet Caravel à l'INRIA).

- assister l'utilisateur dans la mise au point de chaînes de traitement de l'information ("dataflow" scientifique) destinées à produire des informations dérivées utiles. Cette action conduira au développement d'un outil de gestion de chaînes de traitement.
- assurer la documentation et l'archivage des données produites et des programmes de traitement au moyen de métadonnées qui facilitent la localisation et l'identification des ressources partagées. Cette action conduira au développement d'un modèle général de métadonnées et d'un outil de gestion associé. Cet outil permettra l'automatisation de l'inventaire et de la saisie des métadonnées.
- assurer la localisation efficace des informations pertinentes dans le réseau de ressources. Cette action conduira au développement d'un catalogue intelligent permettant d'identifier les données disponibles, de les localiser et d'en appréhender le contenu et la qualité par le biais des métadonnées. Ce catalogue intelligent incorporera des informations sur la localisation, adaptable à divers profils utilisateurs pour les recherches personnalisées.

Les résultats de ces actions seront validés sur deux scenarii d'utilisation de systèmes d'informations environnementaux, qui traitent, pour le premier, de l'observation à long terme des écosystèmes aride et semi-aride pour l'étude et la compréhension des processus de la désertification (Programme ROSELT), pour le second, de la gestion des digues (application SIRS Digues).

3 MÉTHODES DE TRAVAIL

- Les membres du projet se sont réunis aux dates suivantes :
  - Mardi 1er octobre 2002
  - Mercredi 9 octobre 2002
  - Jeudi 27 février 2003
  - Mardi 8 avril 2003
  - Mardi 20 mai 2003
  - Lundi 24 novembre 2003
  - Jeudi 8 janvier 2004
  - Jeudi 11 mars 2004
  - Lundi, mardi, mercredi 7, 8, 9 juin 2004
  - Mercredi 26 janvier 2005
  - Vendredi 8 juillet 2005
  - Mercredi 28 septembre 2005

Des réunions techniques entre certains partenaires ont été organisées tout au long du projet.

- Les transparents présentés lors de ces réunions, ainsi que les comptes-rendus sont disponibles à l'adresse suivante : http://www-poleia.lip6.fr/padoue/documents/index.htm
- Documents produits
  - Rapport d'avancement du projet PADOUE (Janvier 2004) disponible sur http://www-poleia.lip6.fr/padoue/documents/RapportIntermediaire.doc
  - N. Lumineau. Organisation et Localisation de Données Hétérogènes et Distribuées sur un Réseau Pair à Pair. Thèse soutenue en décembre 2005 et disponible sur http://www-poleia.lip6.fr/~lumineau/these/
  - N. Lumineau. Vers un Service d'annuaire de Ressources Réparties et Hétérogènes. Rapport de stage de DEA (Septembre 2002) disponible sur http://www-poleia.lip6.fr/~lumineau/archives
  - J. Tanguy. Modèle de coût pour l'évaluation de stratégies d'auto-organisation par regroupement de nœuds dans les réseaux pair à pair non structurés. Rapport de stage

Rapport de fin de projet retenu dans le cadre de la campagne ACI GRID 2002

de DEA (Septembre 2005) disponible sur http://www-poleia.lip6.fr/~tanguy/files/Rapport_2005.pdf

- N. Lumineau. Tour d'horizon du filtrage collaboratif. Synthèse sur les techniques de filtrage collaboratif (Octobre 2003) disponible sur http://www-poleia.lip6.fr/~lumineau/archives
- B. Granouillac. Développement d'une extension du logiciel SIG ArcGis pour personnaliser la synchronisation, l'édition et l'exportation des métadonnées. Rapport de stage de DESS (septembre 2004) disponible sur http://mdweb.roselt-oss.org/distrib/exportMdweb-rapport.pdf

- Interactions avec les autres projets de l'ACI GRID

Le projet Padoue a suivi de près les travaux menés dans le projet MEDIAGRID, qui traite d'un sujet très proche (médiation de données hétérogènes et réparties), dans un cadre différent (données biologiques). Nous avons assisté à plusieurs présentations concernant ces travaux, dans des conférences, groupes de travail, école de printemps, journées ACI. Par ailleurs, plusieurs membres des deux projets ont participé à l'Action Spécifique *Médiation via les métadonnées* du CNRS en 2003, mise en place et dirigée par T. Libourel, membre du projet PADOUE.
Les travaux concernant la localisation des ressources dans un réseau pair-à-pair ont été présentés dans le workshop Masse de Données en Pair-à-Pair organisé en mars 2005 par le projet MDP2P.
Une application de partage de données et de programme utilisant Le Select et les dataflow a été développée par l'équipe Visages de l'Irisa, dans le cadre du projet ACI Neurobase.

4 RÉALISATIONS

4.1 Conceptuelles

Interopérabilité :

Pour assurer l'interopérabilité des données, nous utilisons le logiciel Le Select, développé par le projet Caravel de l'INRIA. Cet outil permet d'offrir une vue uniforme des ressources, qui restent stockées sur leur site d'origine, dans leur forme d'origine, et de les interroger à l'aide du langage de requêtes SQL. Un des aspects particulièrement intéressant du Select est qu'il permet également l'interopérabilité des programmes. A ce titre, il joue un rôle central dans l'utilisation du modèle de chaînes de traitement scientifiques.
Dans le cadre du projet Padoue, Le Select a été intégré dans une architecture plus générale de partage de ressources (décrite ci-dessous), qui permet d'intégrer des données et des programmes dont la source n'est pas connue à l'avance. Son utilisation a nécessité l'écriture d'adaptateurs pour les données et les programmes considérés dans le projet, et a donné lieu à de nombreuses optimisations améliorant ainsi les performances de cet outil.

Mise au point de chaînes de traitement scientifiques :

Les études scientifiques s'appuient souvent sur des chaînes de traitements de données constituées de programmes de traitements organisés en graphe de flot de données pour analyser et interpréter des données expérimentales. Les avantages qu'apporte l'utilisation de chaînes de traitements par rapport aux programmes monolithiques sont principalement la réutilisation, l'amélioration indépendante de chacun des composants et les opportunités de répartition d'exécution.
Pour aider les utilisateurs à concevoir leurs chaînes de traitement de données, à les exécuter, et à organiser leurs résultats, nous avons défini un langage de dataflow scientifique (DFS), qui permet

de construire graphiquement des chaînes de traitement pour l'analyse de données scientifiques. Pour faciliter l'utilisation de ce langage, nous avons également défini un framework permettant de guider les utilisateurs dans la conception, ainsi qu'un environnement d'exécution, permettant de suivre les exécutions des programmes. La description complète du langage est donnée dans le rapport intermédiaire (http://www-poleia.lip6.fr/padoue/documents/RapportIntermediaire.doc), et son utilisation est illustrée par l'application SIRS Digue.
Le langage DFS et son environnement d'exécution ont été utilisés pour la mise au point de chaînes de traitement scientifiques dans l'application SIRS Digue (voir section 4.3). Nous avons observé un bénéfice en terme d'interopérabilité, qui facilite le couplage et la comparaison de traitements scientifiques.

Structuration et gestion des métadonnées

Structuration autour du standard ISO 19115 geographic metadata

Afin de structurer les métadonnées utilisées pour référencer les jeux de données et permettre de rendre le catalogue ou la référence localisable par une autre application que celle qui l'a produit, les métadonnées sont structurées à partir du standard international de métadonnées pour l'information géographique (ISO 19115). Ce standard peut être vu comme une structure hiérarchique distinguant deux catégories d'éléments : les rubriques (ou sections) et les éléments pouvant être valués.

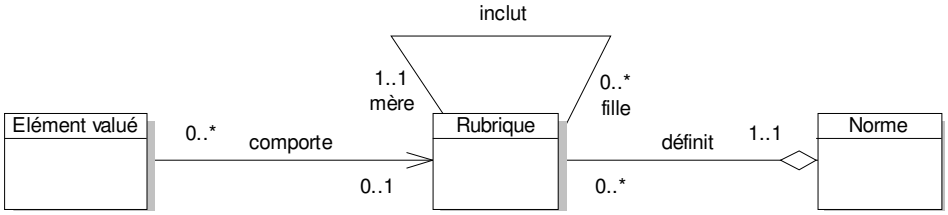


Figure 1 : Modèle hiérarchique du standard International (formalisme UML)

Notion de gabarit

Afin de satisfaire la variété de besoins relatifs aux organismes gérant les métadonnées on introduit la notion de gabarit (dénommé profile par le standard).
Un *gabarit*, ou encore adaptation, est un document et un schéma conceptuel qui précise les options de mise en place du standard afin de répondre à un besoin particulier. Par essence, un gabarit ne contredit pas la norme à laquelle il se réfère. Il décrit plutôt les aménagements que l'on souhaite lui faire subir afin qu'elle puisse être mise en place et utilisée dans un contexte particulier. Les éléments obligatoires du standard doivent bien sûr être respectés mais des éléments n'existant pas dans la norme (éléments étendus) peuvent être intégrés au gabarit. Ils apportent des éléments de description utiles dans le contexte précis pour lequel le gabarit est censé être utilisé. Un gabarit d'une norme permet, en outre, d'adapter culturellement et linguistiquement une norme internationale aux particularités d'un pays ou d'une région.

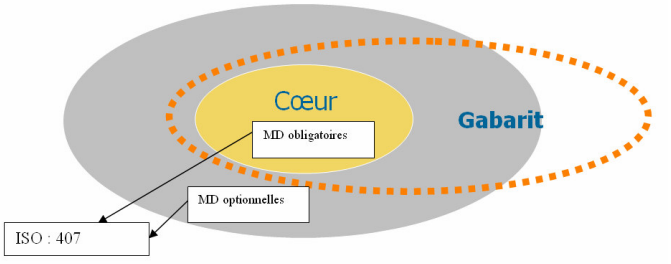


Figure 2 : Représentation schématique d'un gabarit de métadonnées en relation avec le standard international

Une stratification complémentaire (déjà admise par les organismes de normalisation) permet de dégager au sein d'un même gabarit différents niveaux de description et ainsi d'alléger la production de métadonnées.
Cette stratification définit trois niveaux :

- un niveau *élémentaire* correspondant au cœur de la norme
- un niveau *étendu*,
- et un niveau *complet*.

Outil générique
Il nous a semblé nécessaire de proposer aux organismes qui souhaitent mettre en place un service de gestion de métadonnées un outil leur permettant de réaliser ceci de manière souple et simple.
Pour cela il fallait en quelque sorte donner les moyens de générer et d'exporter tout type de gabarit à partir de tout modèle de standard disponible.

Compte tenu des besoins spécifiques de l'application ROSELT, la première réalisation a consisté à construire une base de données (métabase) qui se décompose en trois niveaux :

- schéma de description du standard
- schéma de description des gabarits
- schéma de stockage des métadonnées et des libellés des interfaces homme-machine

C'est à partir des deux premiers niveaux que l'administrateur d'un organisme peut définir et sauvegarder le gabarit nécessaire à ses besoins. La base de métadonnées peut être construite sur le schéma du gabarit. Une fois le gabarit défini, et afin de produire et diffuser les métadonnées via le Web les interfaces de production, de gestion et de consultation des métadonnées ont été construites (en s'appuyant sur le gabarit).

Architecture et choix techniques.

Le choix des composants s'est orienté vers une solution entièrement OpenSource. Sa mise en œuvre s'articule autour de trois composants :

- un système de gestion de base de données (SGBD)
- des composants de dialogue entre serveur HTTPD et SGBD
- un serveur HTTP permettant la publication des données

De manière optimale, MDweb doit être porté sur le SGBD relationnel PostgreSQL (mais d'autres SGBD relationnels peuvent être utilisés). La plate-forme Web doit être supportée par un serveur HTTP Apache sous Unix. C'est le module PHP adjoint au serveur Apache qui assure la communication entre le système de gestion de base de données et le serveur HTTP.

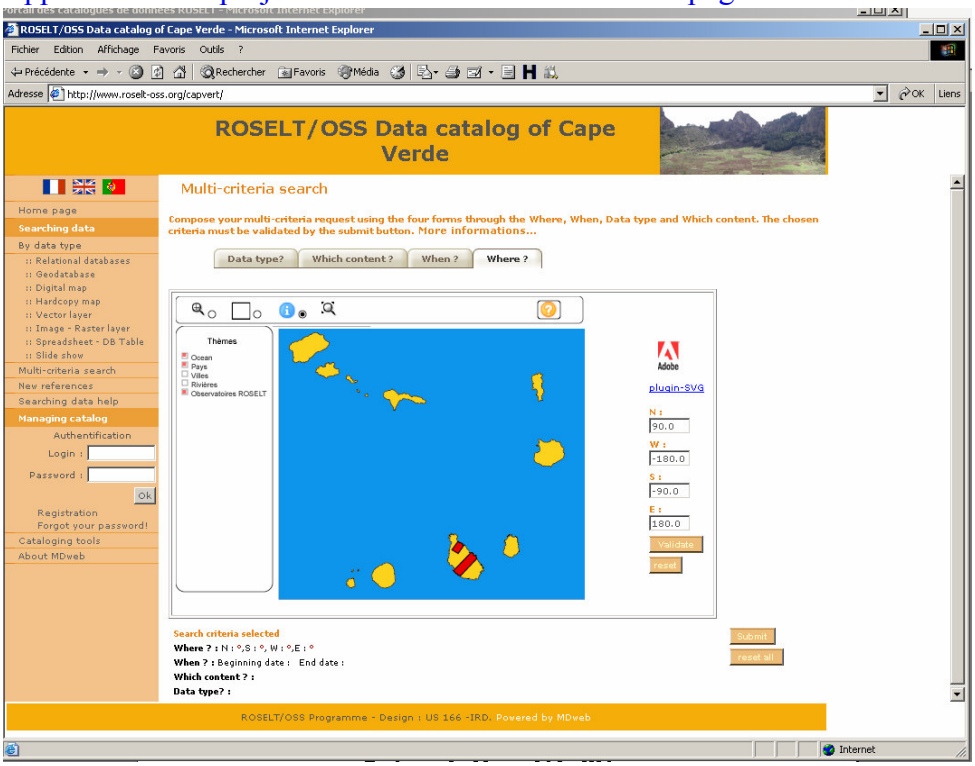


Figure 3 : Interface MDweb

Localisation des informations

L'outil de médiation LeSelect assure l'interopérabilité de données hétérogènes et réparties. Cependant cet outil nécessite la connaissance préalable des adresses des serveurs stockant les données à intégrer. Cette contrainte impose donc que les données soient stockées sur un réseau statique et de taille raisonnable, de manière à permettre aux utilisateurs de se construire une connaissance globale de l'ensemble des serveurs à interroger. Pour tenir compte du contexte de Padoue, où le réseau de ressources peut être amené à se développer et à évoluer fréquemment, et où il n'est plus possible de connaître la localisation des données publiées, nous avons défini une architecture intégrant LeSelect dans une architecture pair-à-pair. Cette architecture permet d'utiliser LeSelect sans avoir à préciser la localisation des données, et permet ainsi le passage à l'échelle du médiateur.

Notre solution permet de spécifier et de stocker toute l'information nécessaire aux médiateurs pour interagir afin de traiter les requêtes initiées par n'importe quel nœud du réseau. Un modèle de schéma sémantiquement enrichi par des informations sur les sources de données a été spécifié. Des schémas de médiation personnalisés, et reflétant la réalité du réseau (les sources disponibles) sont construits dynamiquement, de façon à tenir compte de la large échelle et de l'évolution du réseau. Cette personnalisation résulte d'un étiquetage thématique des données partagées afin de ne fournir aux utilisateurs que les modèles de données concernant leurs thématiques constituant leurs centres d'intérêt. Ainsi, un hydrologue se verra proposer un schéma de médiation modélisant les données sur l'hydrologie réellement disponibles sur le réseau. La construction dynamique des schémas de médiation s'appuyant sur ce principe permet de réduire et de distribuer la charge induite par l'intégration automatique des schémas associés à une même thématique. Cette construction dynamique est basée sur la détection de schémas homogènes (réduction de la charge d'intégration) et de la collaboration des pairs (distribution de la charge d'intégration). Elle a été validée par la plateforme de simulation SimulR, décrite en 4.2.

Afin de localiser efficacement les données dans un tel réseau, nous avons proposé deux approches complémentaires. La première (voir LDD04, LDG04) consiste à organiser le réseau logiquement, selon la sémantique du contenu des nœuds. Cette organisation permet de créer des clusters logiques de nœuds contenant des données de même thématique, ou de thématique proche. Ainsi, une requête émise sur un serveur sera propagée prioritairement vers le (ou les) cluster(s) correspondant à sa thématique. Cette solution s'appuie sur une technique de classification neuronale, adaptée au cadre distribué des réseaux pair-à-pair. L'insertion d'un nœud dans des clusters sémantiques se fait à l'aide d'un service Web (VENISE) décrit en 4.2 interrogeant un classifieur neuronal pour déterminer le voisinage logique le plus pertinent pour ce nœud selon des critères thématiques et physiques. L'originalité de l'approche consiste à exploiter des outils de classification propres à la fouille de données comme les cartes de Kohonen et de proposer un fonctionnement distribué sur plusieurs sites afin d'extraire une connaissance sur la localisation physique du nœud en effectuant la tâche de classification. Le choix technologique des cartes de Kohonen est motivé par l'exploitation de la topologie neuronale pour en déduire une connaissance sur les clusters.

Une deuxième approche pour améliorer l'efficacité de la localisation des données dans le réseau consiste à s'appuyer sur l'expérience des autres membres d'une même communauté pour propager la requête en priorité vers les pairs ayant déjà fourni des données pertinentes aux membres de cette communauté (voir DL03, LD04). Ce routage collaboratif des requêtes repose sur l'existence de deux types de liens sémantiques. Les liens basés sur la pertinence matérialisent la connaissance communautaire en pointant sur les nœuds jugés pertinents par une communauté. Les pairs pointés par ce type de liens contiennent potentiellement des données pertinentes pour le traitement des requêtes. Les liens intercommunautaires permettent de relier deux communautés ayant des centres d'intérêt proches. Cette connaissance permet d'exploiter la connaissance de communautés distantes pour déterminer les pairs pertinents pour traiter les requêtes issues de cette communauté. Les mises en correspondance entre les différentes communautés s'appuient sur une technique de filtrage collaboratif permettant de comparer leurs expériences.

Publications :

[DL03] A. Doucet et N. Lumineau. *A Collaborative Approach for Query Propagation in Peer-to-Peer Systems.* In Proc. of the Semantic Web and Databases Workshop (en marge de VLDB03), Berlin, sept 2003.

[DE03] Desconnets J.C., Moyroud N., Libourel T. : *Méthodologie de mise en place d'observatoires virtuels via les métadonnées*. InforSid 2003, Nancy, Juin 2003.

[LD04] N. Lumineau et A. Doucet. S*haring Communities Experiences for Query Propagation in Peer-to-Peer Systems,* In 8th International Database Engineering and Applications Symposium (IDEAS'04), 7-9 Juillet 2004, Coimbra, Portugal.

[LDG05] N. Lumineau, A.Doucet, S. Gançarski. *Thematic Schemas Building for Mediation-based P2P Architecture,* International Workshop On Database Interoperability (InterDB 2005), Namur, Belgium, April 2005.

[LDD04] Lumineau, N., Doucet, A., Defude, B., Cluster Entries for Semantic Organization of Peer-to-Peer Network, Semantics for Grid Databases (ICSNW'04), Paris, June 2004. [Poster]

[LDG06] N. Lumineau, A.Doucet, S. Gançarski. Distribued Neural Network for Peer Organization*,* Article soumis à ICALT 2006.disponible sur htt://www-poleia.lip6.fr/~lumineau/soumissions/icalt06.pdf

[L05] N. Lumineau Organisation et Localisation de Données Hétérogènes et Distribuées sur unRéseau Pair à Pair. Thèse soutenue en décembre 2005 et disponible sur http://www-poleia.lip6.fr/~lumineau/these/

4.2 Logicielles

Logiciels développés dans le cadre du projet :

- LeSelect : outil de médiation commercialisé par Business Object depuis septembre 2005. (http://www-caravel.inria.fr/~leselect/)
- MDweb : outil d'indexation et de recherche de l'information environnementale (documentation : http://mdweb.roselt-oss.org/accueil/presentMDweb.php , en production : http://mdweb.roselt-oss.org/)
- Export MDweb : extension ArGis pour la synchronisation, l'édition et l'exportation des métadonnées dans le logiciel ArcGis (téléchargement : http://mdweb.roselt-oss.org/distrib/ExportMDweb.zip)
- Dataflow : environnement de définition et d'exécution des chaînes de traitement.
- SimulR (simulateur réparti): plateforme de simulation gérant l'accès à l'ensemble des données associées à un pair simulé par une véritable instance de LeSelect. Le simulateur est distribué sur plusieurs nœuds physiques sur lesquels sont stockés plusieurs pairs logiques. (http://www-poleia.lip6.fr/~lumineau/logiciels/simulr)
- MEnT2 (Mediation in two times) : interface gérant les interactions homme-machine permettant aux fournisseurs de données de configurer leur médiateur, et aux utilisateurs de générer les schémas de gabarits intégrés en fonction des thèmes pertinents. (http://www-poleia.lip6.fr/~lumineau/logiciels/ment2)
- VENISE (Service for node insertion in semantic clusters) : service Web permettant d'interroger un classifieur neuronal, pour l'insertion d'un nœud dans un cluster sémantique adéquat. (http://gaya.lip6.fr:8080/venise)
- ETNA (Experiences Traceability N'Analysis) : interface graphique permettant de parser et d'analyser les journaux générés par SimulR. Cette application permet d'analyser les performances des simulations produites par SimulR. (http://www-poleia.lip6.fr/~lumineau/logiciels/etna)
- MDSearch : interface permettant à un utilisateur d'interroger des données (semi)structurées stockées à travers des Bases de Données dans le réseau. (http://www-poleia.lip6.fr/~lumineau/logiciels/mdsearch)

Démonstrations

- Lumineau, N., Doucet, A., Defude, B., VENISE: Content-based Clustering for Data Sharing in Peer-to-Peer Architecture, BDA'04, Montpellier, Octobre 2004.
- Nicolas Lumineau, Anne Doucet - LIP6, Bruno Defude - INT Evry. Organisation sémantique d'un réseau Pair-à-Pair dédié au partage de données. Journée des Ambassadeurs au LIP6 Paris (http://www.lip6.fr/Direction/2005-06-06.html)
- Nicolas Lumineau, Anne Doucet - LIP6, Bruno Defude - INT Evry. Semantic Organization of P2P network for Data Sharing. Présentation du LIP6 aux représentants des laboratoires Sun (dans le cadre du Pôle de Compétitivité), LIP6. Paris

4.3 Applications

Application « gestion des digues contre les inondations»

Le Ministère de l'Environnement et du Développement Durable (MEDD), avec l'appui du Cemagref, se préoccupe depuis le milieu des années 1990 de la gestion des digues fluviales qui protègent les zones inondables car les enjeux matériels et humains sont considérables.

Deux applications informatiques ont notamment été développées :

- La première application, intitulée BarDigue, fonctionne à l'échelle nationale pour effectuer le recensement des digues afin d'identifier celles présentant le plus d'enjeux et planifier les

investissements. Ce recensement est mis à jour en continu. L'application informatique est hébergée sur un serveur du Cemagref à Aix en Provence. Elle repose sur une architecture classique intranet / SGBD et permet aux organismes chargés d'effectuer localement le recensement des digues de saisir et consulter les données de la base depuis un simple navigateur.

- La deuxième application, intitulée SIRS digues, fonctionne à l'échelle locale. Elle est destinée aux gestionnaires locaux de digues pour gérer leur patrimoine d'information. Elle repose sur le couplage entre un SIG (Système d'Information Géographique) et un SGBD et fonctionne sur des PC autonomes ou en réseau local. Par rapport à l'application nationale, cet outil décrit beaucoup plus finement la structure et l'état des digues et intègre leur représentation cartographique.

Le scénario retenu dans le cadre du projet PADOUE est le suivant :
Le responsable de l'application recensement national balaie régulièrement sur le réseau les nœuds de la communauté SIRS digues pour recenser les tronçons de digue fragiles (fournis par chaque application SIRS digue) puis les croise avec la base de donnée nationale BD Topo de l'IGN dont il dispose sur un autre serveur pour mettre à jour l'inventaire des enjeux (bâtiments sensibles comme les écoles, les hôpitaux, les casernes de pompier, les stations de traitement des eaux) le long de chaque digue.
Ce scénario a ensuite été décrit à l'aide du langage de DataFlow développé par l'INRIA.
Le DFS ou DataFlow Scientifique est un langage permettant d'identifier et de formaliser, au sein d'un programme complexe, des programmes « modulaires » correspondants à leur tour éventuellement à une série de traitements plus élémentaires. Ces programmes modulaires étant identifiés, il est ensuite plus facile de les ré-employer, en particulier dans le cas de l'utilisation de ressources délocalisées avec Le Select. L'usage du DFS peut avoir également l'avantage de servir de support pour conserver la généalogie des données (quels sont les données et les traitements à l'origine ?) et donc d'enrichir les métadonnées.

Le DFS est un langage modulaire dont les pièces élémentaires sont au nombre de trois :

- les « entités de données », c'est à dire les données en entrée ou en sortie des traitements ;
- les « dataflow units » (DFUs) sont des chaînes de traitement élémentaires prenant en entrée ou produisant en sortie les « entités » précédentes ;
- les datalinks spécifient les flots de données entre les DFU via des entités de données.

Les DFUs peuvent désigner une chaîne de traitements complexe elle-même composée d'autres DFUs liés entre eux par des datalinks. Un dataflow uni peut donc être une vue « encapsulée » d'autres DFUs.

Le scénario décrit verbalement à la page précédente peut être représenté graphiquement sous la forme du dataflow suivant (figure 4) :

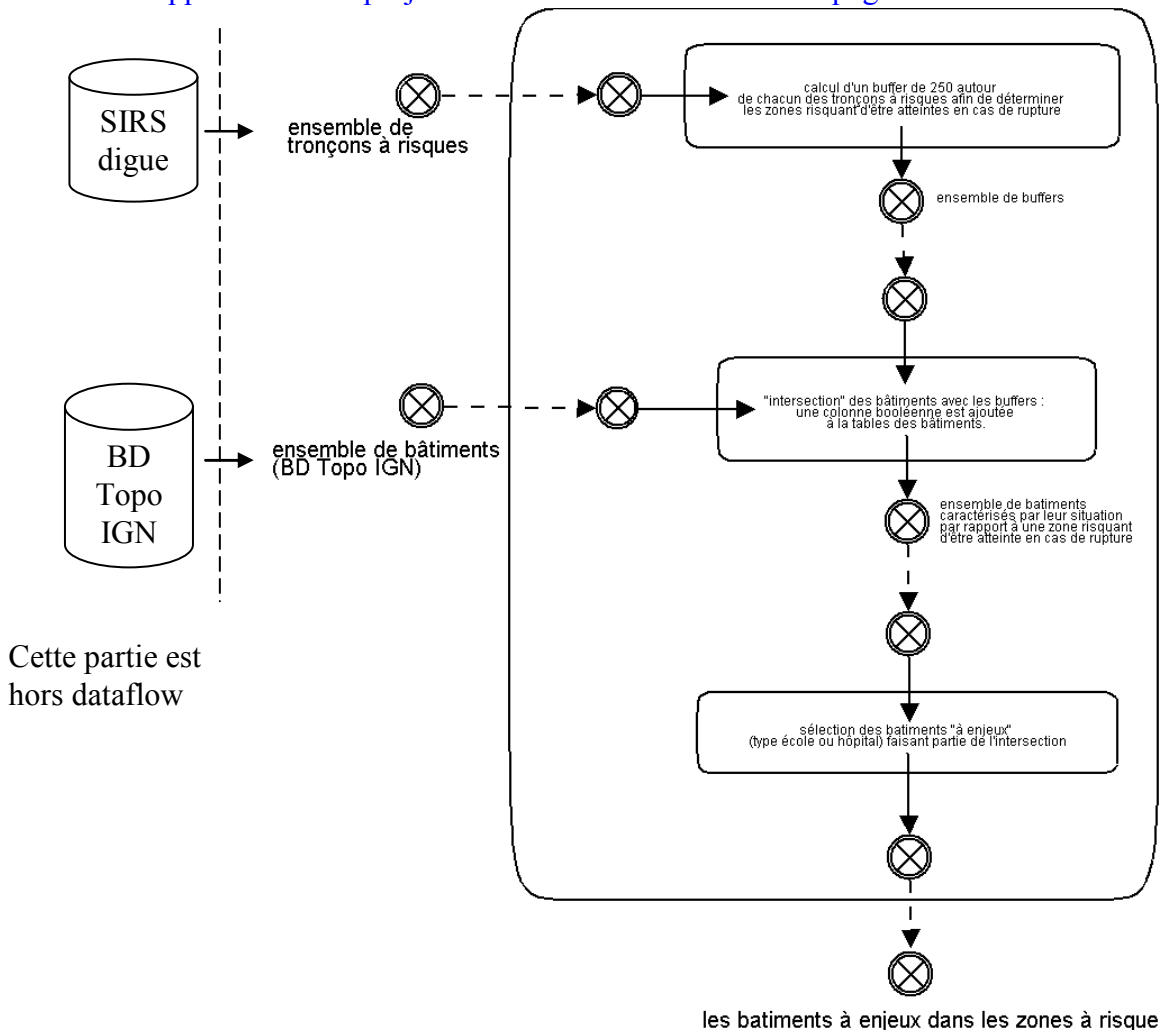
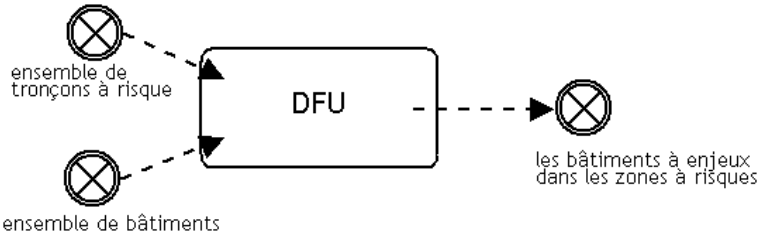


Figure4 : Représentation graphique du dataflow

- Les ⊗ sont des entités de données du type « ensemble » ;
- Les - - - - → et les ——→ sont les datalinks ;
- Une autre représentation « encapsulée » du dataflow mettant en valeur l'élément DFU est la suivante :



A l'intérieur de cette DFU, d'autres DFUs plus élémentaires peuvent être distinguées.

Les diagrammes de DFS sont réalisés grâce à une bibliothèque de symboles spécifique rajoutée à DIA (utilitaire gratuit pour la création de diagrammes et schémas divers).

La mise en œuvre du dataflow

Elle aurait dû être automatique via l'utilisation des fonctionnalités de Le Select. Cependant, la commercialisation de Le Select a limité la portée du test.
Pour la réalisation pratique du test, le SGBD Postgres et son extension spatiale Postgis ont été utilisés. Les données en entrée et les données créées en sortie sont toutes stockées dans une même base de données. Un programme Java a été élaboré : il prend en entrée les données disponibles dans cette base, leur applique des requêtes exprimées en langage SQL standard et étendu correspondant aux DFUs élémentaires. Le résultat de l'exécution du programme est une table contenant la géométrie des bâtiments à enjeux.

La visualisation a pu se faire grâce au logiciel SIG JUMP 1.1.0. (figure 5). Ce logiciel assez rudimentaire donne la possibilité de visualiser et de réaliser des traitements sur des données géographiques, notamment celles stockées dans des bases PostGis.

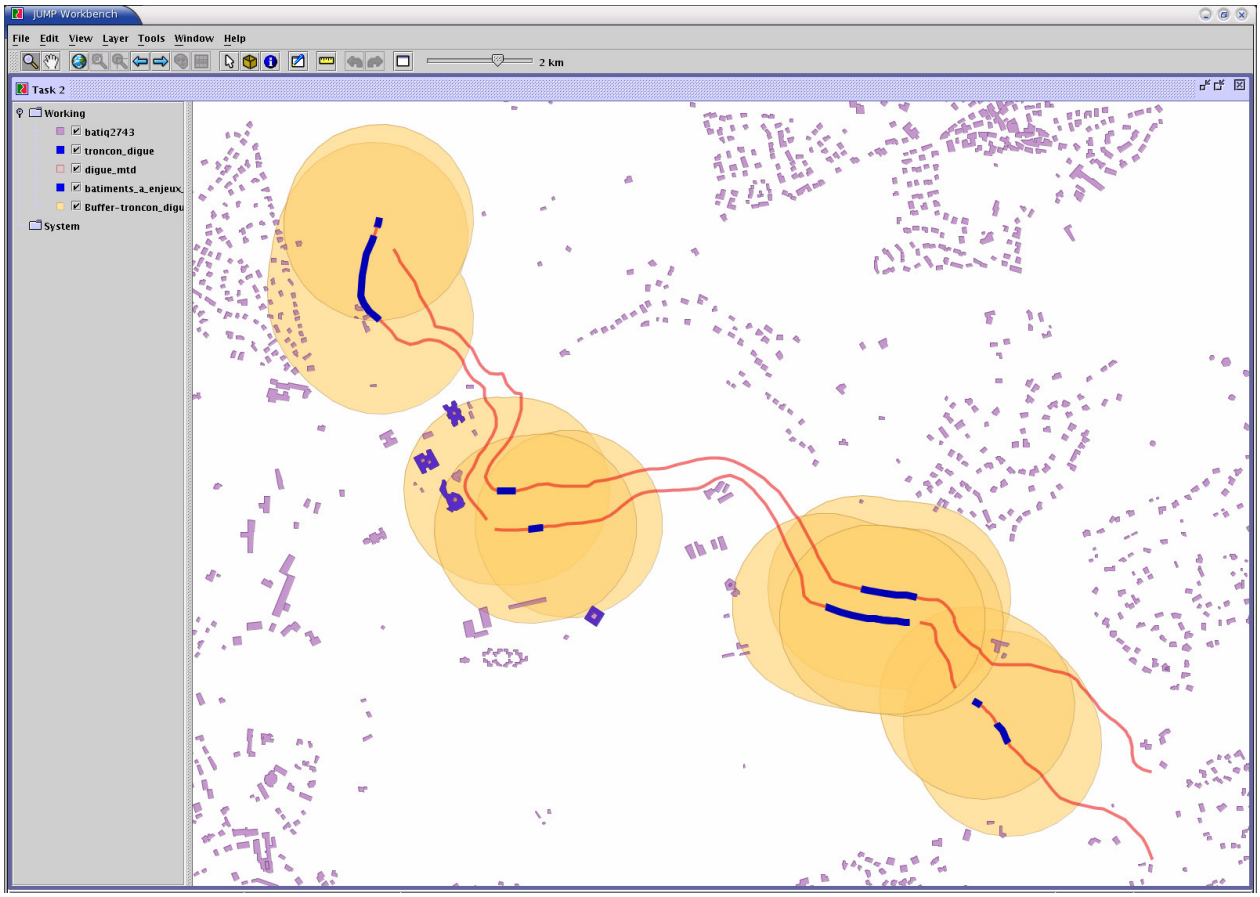


Figure 5 : Illustration du résultat issu du dataflow (orange : digues ; bleu : tronçons de digue fragiles ; violet pâle : bâtiments BD Topo IGN ; jaune : zones potentiellement impactées par une rupture de digue ; violet foncé : bâtiments à enjeux)

Action Concertée Incitative
[ACI]
Globalisation des Ressources
Informatiques et des Données
[GRID]

Rapport de fin de projet retenu dans le cadre de la campagne ACI GRID 2002

Application de l'outil d'indexation et de recherche de l'information environnementale MDweb dans ROSELT

MDweb est au cœur du système de circulation de l'information environnementale du programme ROSELT (http://www.roselt-oss.org/). L'objectif de ce système est double, il s'agit :
- de porter à la connaissance de tous les données produites par les observatoires (collectées et traitées) par l'élaboration d'une documentation standardisée accessible à tous et facilement consultable,
- de mettre à disposition une information scientifique comparable, de qualité et cohérente dans le temps et dans l'espace, pour sa valorisation dans le cadre de la lutte contre la désertification dans le respect du droit, de la propriété intellectuelle et de la confidentialité tels qu'ils seront définis par les membres du programme ROSELT.

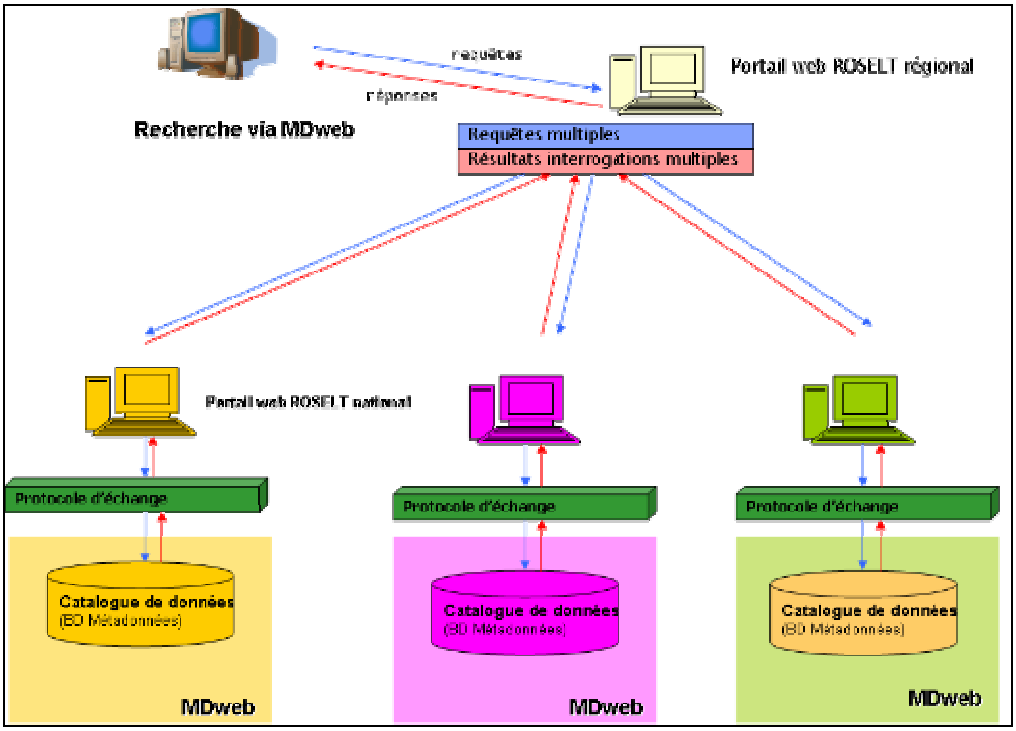


Figure 5 : Architecture de la circulation de l'information dans le programme ROSELT au niveau régional

L'élaboration d'un tel service au sein du réseau d'observatoires est née de la nécessité de constituer un patrimoine d'informations documentées et facilement accessibles qui puissent être réutilisées dans le cadre d'études actuelles ou futures hors du contexte de production de l'information. Elle répond, de manière générale aux besoins des écologues et des scientifiques de l'environnement de pouvoir se doter de séries temporelles suffisamment longues sur lesquelles ils puissent appuyer des études diachroniques utiles à l'analyse des changements environnementaux.

Conformément à l'organisation du réseau d'observatoires et aux missions qui sont attribuées aux institutions nationales en charge de la surveillance environnementale, MDweb est installé dans chaque institution nationale, généralement sur le serveur web, et intégré au site institutionnel lorsqu'il en existe un. Ce sont les membres de l'équipe ROSELT du pays concerné, appuyés par le chargé de Système d'Information, qui ont pour tâche de constituer et de gérer leur propre catalogue de données dans MDweb. A terme, chaque pays du réseau ROSELT possédera un catalogue de données consultable en ligne via son site institutionnel.

L'ajout d'un module de connexion Z39.50 à MDweb permettra dans un deuxième temps de proposer une recherche et consultation sur l'ensemble des catalogues ROSELT à partir du portail régional du programme.

5 RETOMBEES ESSENTIELLES

Les travaux menés dans le cadre du projet Padoue ont permis d'aborder plusieurs nouvelles recherches, essentiellement dans le domaine de la gestion de métadonnées et du partage de ressources hétérogènes et distribuées, dans les réseaux à large échelle. Ils ont également eu des retombées significatives pour le CDS, pour le Cemagref et l'IRD.

Retombées pour les laboratoires universitaires

L'équipe Bases de Données du LIP6 participe actuellement à plusieurs projets dans le domaine de la gestion de données dans les réseaux pair-à-pair :

- **Projet LIP6** : raffinement sémantique du processus de localisation de ressources par profil utilisateur
- **Projet SemWeb** (ACI Masse de données 2004) : interrogation du Web sémantique avec XQuery (http://bat710.univ-lyon1.fr/~semweb/)
- **Projet RARE** (projet incitatif GET) : routage par apprentissage de requêtes (http://www-inf.int-evry.fr/~defude/RARE/)

Elle est également porteur du **projet RESPIRE** (ARA Masse de données 2005) : ressources et services pair-à-pair, interrogation et réplication.

Les travaux menés dans le cadre de Padoue et les compétences que nous avons acquises dans le domaine de partage de données à large échelle nous ont permis de participer au projet Infomagic pôle de compétitivité IMVN (Image, Multimedia et Vie Numérique).

Une thèse portant sur le regroupement de nœuds dans les architectures pair-à-pair, et financée par la région Ile de France, a commencé en novembre 2005. Elle se situe tout à fait dans la continuité de travaux effectués dans ce domaine dans Padoue.

Le LIRMM et les membres concernés de l'équipe DOC (Données Objets Composants pour les systèmes complexes) étaient déjà impliqués dans le réseau de recherche régional (Cemagref, Cirad, Ird) axé sur la gestion et le traitement de l'information géographique par l'intermédiaire de séminaires organisés dans le cadre du GDR Cassini dans lequel nous assurions la responsabilité de l'axe Qualité et Métadonnées. Les enjeux de partage et mutualisation mis en lumière au travers du projet Padoue ont largement essaimé et permis de contribuer à des projets connexes sur la gestion intégrée (thèse de Julien Barde), sur l'intégration de bases de données géographiques (collaboration avec le laboratoire COGIT-IGN), sur la médiation de bases de données génomiques (thèse de P Larmande) et l'intégration de bases de données biologiques (ACI ISIbio). Les réflexions menées dans le contexte Padoue ont révélé le rôle essentiel de la sémantique afférente aux domaines pluridisciplinaires des sciences de l'environnement. La construction et révision d'ontologies (a priori par interaction directe des utilisateurs) ont déjà été intégrées, nous souhaitons prolonger cette approche par une démarche plus théorique basée sur l'analyse formelle de concepts.

Les retombées se situent au niveau de la formation (master pro et recherche) via la proposition de nouvelles unités autour de la médiation (avec l'outil Le Select), de la gestion de métadonnées et d'ontologies et par la participation active de stagiaires. Sur le plan expertise, la capitalisation des approches tant conceptuelles qu'opérationnelles nous a permis d'être sollicités comme experts dans le contexte de divers projets environnementaux (Projet Observatoires Cirad, Projet COPT – ADD

2006, Projet Diren), de resserrer les liens avec le réseau Géoïde (Canada) et d'être porteur d'un projet de recherche (PPF) déposé dans le cadre du plan quadriennal de l'Université Montpellier II. Enfin, l'ensemble des travaux menés nous permet de participer au Drafting Team Inspire sur les groupes spécifications des données et métadonnées.

Retombées pour le CDS :

Dans le cadre du projet PADOUE, le CDS avait pour objectif d'évaluer l'impact potentiel des différents travaux sur certains de ses services. Le CDS maintient plusieurs services de référence utilisés par la communauté astronomique internationale. On peut citer par exemple Aladin, Simbad et VizieR pour lesquels le nombre d'accès quotidiens se compte en dizaines de milliers. Un service de catalogues comme VizieR offre un accès à environ 5000 catalogues dont certains dépassent le milliard de lignes. La masse de données ne cesse de croître et il est vital pour le CDS d'explorer les possibilités offertes par les nouvelles technologies, et d'exploiter celles qui semblent les plus pérennes.

Les apports du projet PADOUE concernent plus particulièrement l'outil de médiation LeSelect, les chaînes de traitement scientifiques et les réseaux pair-à-pair.

Le CDS a bénéficié d'une formation au Select en 2004, qui peut être utilisé, dans le cadre des travaux menés par l'IVOA (International Virtual Observatory Aliance), pour accéder aux sources de données via SQL, et permettre d'assurer l'interopérabilité de données stockées dans des formats hétérogènes. Le Select peut aussi être utilisé par le service VizieR, qui n'offre pas de possibilité d'interrogation via SQL actuellement. Le Select permet d'offrir un accès SQL sans être obligé de stocker les grands catalogues dans des bases de données relationnelles. Le service VizieR peut alors être interrogé par des requêtes SQL quelque soit le catalogue concerné, cette opération devenant complètement transparente.

Un autre objectif du projet PADOUE était d'assister les utilisateurs dans la mise au point de chaînes de traitements. Les notions de Dataflow et de Workflow ont été exposées par Jean-Pierre Matsumoto, en mars 2004. Depuis, un groupe de travail interne au CDS travaille sur l'utilisation des workflows au niveau du CDS. Ceux-ci sont maintenant également exploités au niveau du projet Masses de Données en Astronomie (Architecture AÏDA, Astronomical Image processIng Architecture) de l'ACI Masses de Données. Par ailleurs un groupe de travail consacré aux workflows s'est constitué au niveau de l'Observatoire Virtuel français.

Les travaux concernant la localisation des métadonnées offrent des perspectives intéressantes au niveau des registres développés actuellement par les membres du groupe de travail « Registry » de l'IVOA. La mise en œuvre d'un registre unique, même répliqué, n'est pas envisagée pour de multiples raisons et il convient de mettre en œuvre des outils permettant d'accéder à des registres distribués, qui contiendront les métadonnées les plus pertinentes. Le travail réalisé par N. Lumineau devrait apporter des solutions à certains problèmes non résolus à ce jour.

Retombées pour le Cemagref :

Le Cemagref s'est intéressé au partage de données hétérogènes réparties dans le cadre de réflexions menées notamment au sein du réseau d'animation interne Reglis (Représentation et Gestion de l'information spatialisée), piloté au départ par P.Maurel avec la participation de G.Bonnet. Il s'agissait pour l'établissement, et sa Direction Scientifique en particulier, de s'appuyer sur les recherches dans le domaine de la médiation des données pour identifier des solutions permettant d'inventorier les très nombreuses données disponibles dans les neuf sites du Cemagref, puis éventuellement les partager ou les intégrer dans des applications complexes. Les premières approches faites avant le démarrage de Padoue avaient ouvert des pistes intéressantes pour le couplage de données hétérogènes et distantes et le développement d'applications intégrant ces données.

Compte-tenu de ce contexte et des partenariats déjà existants avec le Lip6, le LIRMM et l'INRIA, le Cemagref s'est naturellement impliqué dans le projet PADOUE pour participer au développement et à l'expérimentation de nouvelles solutions pour la médiation des données, l'objectif à terme étant de pouvoir arriver à des solutions opérationnelles.
La réalisation du projet PADOUE a tout d'abord permis de renforcer les partenariats avec le LIP6, le LIRMM, l'INRIA et l'IRD, notamment par le biais d'échanges scientifiques autour de la thèse de Nicolas Lumineau, et en externe au projet, de la thèse de Julien Barde centrée sur les métadonnées. Sur le plan scientifique, ce projet a été l'occasion d'approfondir au niveau conceptuel l'articulation entre métadonnées, langage de dataflow et médiateur au sein d'une architecture de médiation. L'intérêt d'une approche P2P pour des applications environnementales a également pu être largement débattue et évaluée. Sur le plan opérationnel, l'outil Le Select n'a pas pu être pleinement exploité pour diverses raisons. Ceci a réduit la portée des tests sur la médiation des données et l'exploitation du dataflow au profit d'un approfondissement de la composante métadonnée.
Dans le courant de l'année 2005 et en parallèle au projet PADOUE, le Cemagref a mené une enquête interne pour recenser des bases de données et évaluer leur contenu et les modalités de gestion. Il ressort que les priorités devront porter, tout d'abord sur la mutualisation des données dites de référence (IGN, INSEE, orthophotos, ...), puis sur la constitution et la mise à jour de bases de métadonnées. Un des enjeux, approfondi dans la thèse de Julien Barde, portera sur le développement de référentiels sémantiques pour mieux contrôler l'indexation des métadonnées et améliorer ainsi la recherche des données pertinentes.
A moyen-terme, l'outil de médiation (Le Select ou un autre) restera une composante fondamentale de l'architecture envisagée pour accéder à des données réparties. Cet outil pourrait également servir à la médiation entre plusieurs bases de métadonnées, la solution testée jusqu'à présent reposant sur une base de métadonnées unique et centralisée. Ce principe est d'ailleurs déjà utilisé dans l'outil de métadonnées MDWeb de l'IRD en exploitant la norme Z39.50. Les tests devront donc être poursuivis à l'avenir.
A plus long terme, le langage de dataflow utilisé dans Padoue nous paraît être une piste prometteuse pour plusieurs raisons : Il permet de décrire des traitements mobilisant des données distribuées. Il permet aussi de décomposer des traitements complexes en sous-ensembles ce qui devrait a priori faciliter leur réutilisation dans d'autres applications. Le dataflow lui-même, avec éventuellement les codes de calcul correspondants, peut être enregistré dans la rubrique « généalogie » des métadonnées qui décrivent les données obtenues à l'issue des traitements. Enfin, ceci autorise des exécutions ultérieures de la même chaîne de traitement, soit sur le même jeu de données d'entrée, soit sur d'autres jeux de données.
Ce dernier à travers les avancées sur la conception des métadonnées, les relations à établir entre les sites Cemagref et à travers un outil de médiation (Le Select ou un autre) devrait nous permettre d'avancer sur notre projet d'inventaire et de partage de données. Quant au dataflow, plus complexe à mettre en œuvre, son utilisation nécessitera d'autres recherches.

Retombées pour l'IRD

C'est avant tout dans le cadre du programme ROSELT que l'IRD a pu bénéficier des recherches conceptuelles et logicielles réalisées dans le cadre de PADOUE. Les avancées conceptuelles et les réalisations logicielles, dont le projet PADOUE a été pour une bonne part le catalyseur, ont pu donner à l'IRD et notamment l'US Désertification, une expertise dans la structuration des métadonnées et leur gestion et la doter d'un outil générique : MDweb. Mise en œuvre chez les différents partenaires du Sud (Egypte, Université d'Alexandrie ; Tunisie, Institut des Régions Arides ; Sénégal, Centre de Suivi Ecologique ; Cap vert, Institut National de Recherche et Développement Agraires, etc.), l'approche menée a fait émerger un besoin plus large au sein des organismes africains nationaux et sous-régionaux. Au sein de l'IRD, l'outil et l'expertise

développée seront aussi valorisés en mettant à disposition l'outil dans d'autres équipes de recherche et au niveau central pour servir de moteur de stockage des catalogues de données géographiques de l'institut. Dans un proche avenir, l'IRD souhaiterait utiliser la démarche afin de réaliser l'inter-connexion des catalogues de données environnementales de l'institut avec ses partenaires du Nord et Sud.

6 UTILISATIONS DU BUDGET

	Matériel	Fonctionnement	Missions	Salaires	**Total (€ HT)**
LIP6		6688	28428	7625	42741
INRIA			1409	108675	110084
Cemagref	2190	5978	5250	7133	20551
LIRMM	11302	4143	9151	17053	41649
IRD	2000	10536.8			12536.8
Total	15492	27345.8	44238	140486	227561.8

Le LIP6 a pris en charge toutes les missions du CDS, pendant la durée totale du projet.

Une bourse de thèse du ministère a été accordée au projet PADOUE, et a permis de financer la thèse de Nicolas Lumineau, au LIP6. Ce dernier a travaillé sur la localisation des ressources dans le réseau. Dans le cadre du projet, il a proposé une architecture de médiation pair-à-pair, et deux approches complémentaires de localisation des données. Il a également réalisé plusieurs prototypes pour valider ses propositions. Sa thèse a été soutenue le 5 décembre 2005. L'obtention de cette bourse a été cruciale pour le bon déroulement de cette action de recherche dans le projet.

7 SUGGESTIONS D'AMELIORATION

Comme mentionné dans ce rapport, l'ACI GRID a eu un impact largement positif sur les membres du projet Padoue et sur leurs organismes. Nous sommes donc tous très favorable à une poursuite de ce type d'actions, notamment sur les moyens (architectures, modèles, outils) permettant la mise en œuvre d'application mettant en jeu de grands volumes de données aux caractéristiques diverses (données environnementales, données spatio-temporelles, données provenant de capteurs, etc.).
L'aspect pluridisciplinaire a constitué un point fort du projet, tant pour les informaticiens que pour les chercheurs des organismes (Cemagref, IRD, CDS). Nous pensons qu'il est très important de pouvoir confronter les avancées en recherche à des problématiques et applications concrètes. Nous souhaitons donc fortement voir se développer de nouveaux projets pluridisciplinaires.
Une difficulté à laquelle sont confrontés les chercheurs est le financement de doctorants et post-doctorants. L'ACI-GRID nous a permis de bénéficier d'un financement de thèse, qui a été très profitable. Pour mener à bien les projets de recherche, il est indispensable d'augmenter et de faciliter ce type de financement.