32nd IAMSLIC Annual Conference
Every Continent, Every Ocean
October 8-12, 2006
Portland, Oregon, U.S.A

Fred Merceur

Mise en place, à l'Ifremer, d'une Archive Institutionnelle et d'un moissonneur OAI spécialisé en sciences marines et aquatiques

Résumé

En Août 2005, l'Ifremer, lançait Archimer, son archive institutionnelle : une base de textes intégraux permettant d'accéder gratuitement à un ensemble de publications, de thèses, d'actes de congrès et de rapports internes. S'inscrivant dans le cadre du mouvement Open Access, cette base contribue à valoriser les travaux de l'Ifremer au plan international. Une année après son ouverture Archimer propose plus de 1400 documents dont plus de 70% des publications rédigées ou co-rédigées par l'Ifremer depuis août 2005.

Dans sa démarche de soutient au mouvement Open Access, l'Ifremer, par le biais de la Bibliothèque La Pérouse, a également développé Avano, un moissonneur OAI spécialisé en sciences marines et aquatiques. Ce moissonneur permet de collecter les données bibliographiques des ressources électroniques (documentation, images, données brutes, vidéos, fichiers audio) stockées dans un ensemble d'Archives Ouvertes et de les agréger dans une base de données centralisée. Ce moissonneur permet non seulement d'indexer les références des ressources issues des archives des organismes de recherches spécialisées en sciences marines, mais également une sélection de ressources, liées aux sciences marines, déposées dans d'autres archives ouvertes (ex : ArXiv, PubmedCentral, …).

Mots-clés: Libre Accès, Archive institutionnelle, Archive Ouverte, Moissonneur OAI, Post-publication, Archimer, Avano, Documentation électronique

1. Introduction

Depuis le début des années 90, afin de contrer les politiques commerciales abusives de certains éditeurs scientifiques, des communautés scientifiques ont créé des serveurs de pré-publications pour offrir un accès gratuit et immédiat à leurs travaux (ex : ArXiv, en physique et RePec, en économie).

En 2001, l'organisation OAI (Open Archive Initiative) a formalisé un protocole d'interrogation de ces archives. Le protocole OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) a pour but de permettre l'interopérabilité des Archives Ouvertes. En effet, si les Archives ne pouvaient pas communiquer entre elles, un utilisateur, pour trouver un document, devrait interroger les archives l'une après l'autre. Devant la multiplication des projets d'archives, il devient aujourd'hui impossible de mener une recherche efficace par cette méthode.

Pour simplifier l'accès à la documentation disponible dans les archives, le protocole OAI-PMH définit deux rôles :

- Les **fournisseurs de données** "data providers" créent des archives, offrant ainsi un accès aux ressources qu'ils y enregistrent. Les archives compatibles OAI-PMH, offrent la possibilité de collecter (ou de moissonner) les données bibliographiques de leurs ressources par l'intermédiaire d'une série de commandes standardisées définies dans le protocole OAI-PMH.

- Les **fournisseurs de service** peuvent venir collecter les données bibliographiques de plusieurs archives et les rassembler dans le but de créer leur propre base de données. Ils peuvent ainsi ouvrir à leurs usagers, la possibilité d'interroger des bases de données correspondant à la totalité ou à une partie de plusieurs archives. La base Oaister, par exemple, indexe la totalité de plus de 700 archives. Les notices de ces bases de données proposent finalement des liens hypertextes vers le texte intégral des documents, qui eux, restent hébergés sur les serveurs des archives.

Avec Archimer, l'Ifremer se situe donc, dans le cadre du mouvement Open Access, comme un fournisseur de données. Avec le développement d'un moissonneur spécialisé en sciences marines, l'Ifremer se présente également comme un fournisseur de service.

2. Archimer, archive institutionnelle de l'Ifremer

2.1. Intérêt pour l'Ifremer

2.1.1. Soutenir le mouvement Open Access

L'ouverture d'une archive institutionnelle est un soutien concret au mouvement « Open Access », dont l'Ifremer pourrait, à long terme, profiter des avancées. En effet, depuis plusieurs années l'Ifremer subit, de la part des plus grand éditeurs scientifiques, et comme toutes les autres grandes bibliothèques scientifiques, des augmentations de coûts d'abonnements aux revues scientifiques sans aucun rapport avec l'inflation. Ces augmentations l'obligent à consacrer une partie toujours plus importante de son budget aux contrats d'abonnement aux revues, et ce, au détriment d'autres sources d'information.

Si la majorité des publications de l'ensemble de la communauté scientifique internationale devenait accessible gratuitement sur le WEB, à travers un réseau d'Archives Ouvertes, elles pourraient constituer une réelle alternative aux abonnements proposés par les éditeurs scientifiques. Même sans imaginer pouvoir un jour nous passer de ces abonnements, nous pouvons envisager être à terme mieux armés pour négocier, du fait de cette nouvelle donne, nos contrats avec les grands éditeurs scientifiques.

2.1.2. Valoriser la production scientifique

Si l'accès gratuit à l'ensemble de la documentation scientifique internationale est un objectif à long terme, la mise en place d'une archive institutionnelle à l'Ifremer devrait pouvoir avoir un effet immédiat sur la visibilité de ses travaux. En effet, plusieurs études démontrent que les articles en libre accès sont plus cités que les articles uniquement accessibles à partir des sites WEB des éditeurs scientifiques (voir réf. 1,

2, 3, 4). La diffusion gratuite des publications de l'Ifremer, via Archimer, pourrait donc améliorer sensiblement leur impact scientifique.

2.1.3. Créer une nouvelle base de données dédiée aux sciences de la mer

Lorsque le nombre de documents disponibles dans Archimer aura atteint une masse critique, nous espérons que le personnel de l'Ifremer considérera cette base, non seulement comme un moyen de valoriser ses travaux à l'extérieur de l'Institut, mais aussi comme une **base de travail utile à ses recherches**. **Cette base devrait, en effet, agréger un ensemble de documents aujourd'hui disséminés sur plusieurs serveurs**. Elle devrait également donner accès à des documents, et, notamment, à des thèses, auxquelles Archimer est actuellement le seul moyen d'accès.

2.1.4. Renouer des liens entre les équipes de recherche et les bibliothèques

Les équipes de recherche utilisent aujourd'hui massivement les ressources électroniques (bases de données bibliographiques, journaux électroniques…) mises à leur disposition par les bibliothèques. Le personnel de l'institut ne se déplace donc plus, ou peu, dans les salles de lecture des bibliothèques. Il a, en effet, accès à toutes ces ressources directement à partir de son poste de travail.

Cette situation constitue, bien sûr, un réel progrès. Elle permet, par exemple, à tout le personnel de l'institut, quelle que soit sa localisation, d'avoir un accès à une très grande partie de la documentation mise à disposition par les bibliothèques et de bénéficier d'outils performants de recherche et de veille documentaires (ex. : bases de données bibliographiques, alertes de recherches automatiques…).

Cette situation a, par contre, tendance à isoler le personnel des bibliothèques des équipes de recherche. Ces dernières peuvent ainsi être amenées à mésestimer le travail effectué par les bibliothèques, dont l'élaboration et la mise en œuvre d'une politique d'acquisition des droits d'accès à des sources d'information sélectionnées). Nous rencontrons, par exemple, régulièrement des chercheurs qui pensent que les articles des éditeurs scientifiques sont accessibles gratuitement sur le WEB. En effet, comme les accès aux ressources de ses éditeurs sont protégés par contrôle de l'adresse IP, les chercheurs y accèdent de manière transparente, sans imaginer, par exemple, l'importance du travail que nécessite la négociation d'un contrat d'abonnement avec un éditeur comme Elsevier.

La mise en place d'une archive offre l'occasion aux personnels des bibliothèques de renforcer ses contacts avec les chercheurs, par l'intermédiaire, par exemple, de la collecte personnalisée des publications à enregistrer dans Archimer.

2.1.5. Améliorer la visibilité du site Internet de l'Ifremer

Les documents enregistrés dans l'Archive de l'Ifremer sont, non seulement, accessibles par l'intermédiaire du site WEB de consultation d'Archimer, mais aussi par l'intermédiaire de moteurs de recherche et par celui de moissonneurs OAI-PMH.

L'étude des statistiques d'utilisation d'Archimer démontre que les moteurs de recherche, dont, principalement Google, sont les principaux points d'accès à nos documents. Elle met en évidence le fait qu'une partie des utilisateurs, qui accède à nos documents directement par un moteur de recherche, continue ensuite sa visite vers la page d'accueil de notre site WEB. A partir de cette page, certains d'entre consultent d'autres documents disponibles via ce site. D'autres utilisateurs, quant à eux, continuent leur visite vers le site WEB institutionnel de l'Ifremer et y découvrent d'autres informations relatives à l'Institut.

Les documents enregistrés dans Archimer, sont, par conséquent, autant de nouveaux points d'entrée au site WEB institutionnel de l'Ifremer. Ils contribuent donc à améliorer l'audience du site WEB de l'Ifremer : site qui est un outil de communication essentiel pour l'institut.

2.2. Principes généraux

2.2.1. Choix de la plateforme de développement

Pour réaliser le système Archimer, nous avons fait le choix interne à l'aide des technologies Java, JSP et Oracle, car nous avions, à l'origine du projet, l'objectif de réutiliser une partie des modules d'Archimer

dans le cadre d'autres projets de la bibliothèque et, notamment, celui d'un projet de rénovation des sites WEB de consultation de nos catalogues. Nous souhaitions, de plus, lier ce nouveau système à d'autres modules informatiques déjà existants comme notre base « Bibliométrie » ou notre portail de revues électroniques. Un développement spécifique nous semblait alors la meilleure solution pour atteindre nos objectifs.

Les sites WEB de ce projet sont développés à l'aide de pages JSP et sont exécutés par le serveur Apache/Tomcat central de l'Ifremer. Ces technologies ont été sélectionnées pour leur conformité avec la politique générale du département Informatique de l'Ifremer.

Les données bibliographiques des documents enregistrées dans Archimer sont stockées dans une base de données Oracle : base hébergée sur un serveur mutualisé entre tous les départements de l'Ifremer. L'utilisation d'Oracle est particulièrement intéressante dans le cadre d'Archimer car elle permet l'intégration de fonctionnalités de recherche documentaire avancées.

2.2.2. Types de documents enregistrés

Actuellement Archimer a été conçue pour permettre l'enregistrement et la diffusion de thèses co-financées par l'Ifremer, de rapports internes, d'actes de congrès et d'articles publiés dans des journaux scientifiques.

Pour faciliter l'acceptation de ce projet par le personnel de l'Ifremer, nous avons souhaité limiter, dans un premier temps, l'enregistrement des articles aux seules post-publications. Leur diffusion gratuite sur le WEB fait aujourd'hui l'unanimité, contrairement à la diffusion des pré-publications qui est parfois critiquées par certains auteurs par peur de plagiat, ou car la qualité de la publication n'est pas assurée par une validation par les pairs.

2.2.3. Format de diffusion

Nous avons choisi le format PDF comme format unique de diffusion. Tous les documents enregistrés dans Archimer sont donc convertis en PDF, et ce quelque soit l'outil utilisé pour sa rédaction (Word, Latex…). Nous avons sélectionné ce format de diffusion pour les raisons listées ci-dessous :

- L'assurance de la pérennité du format PDF, du fait de sa très large utilisation et de la publication de ses spécifications,
- Sa mise en œuvre simple, qui permet de réduire le temps de traitement et d'enregistrement des documents dans Archimer,
- Sa bonne adaptation à la diffusion électronique de documents volumineux, comme les publications ou les thèses.

2.2.4. Conservation des documents

La conservation des documents à très long terme n'a pas, jusqu'à présent, été l'une de nos préoccupations majeures dans le cadre de la définition de ce projet. D'emblée, nous avons, par exemple, exclu de convertir les documents en XML/SGML pour s'assurer de leur pérennité. Le temps de traitement d'une telle conversion nous semblait, en effet, incompatible avec les ressources humaines mises à disposition du projet.

Toutefois, considérant la masse de documents aujourd'hui stockés en PDF, nous espérons que, si ce format devenait un jour obsolète, il existerait alors des outils de conversion de nos fichiers PDF qui nous permettraient de convertir facilement ces fichiers PDF.

2.2.5. Modalités d'enregistrement des documents

Les documents sont enregistrés dans Archimer par le personnel des Bibliothèques de l'Institut, qui assure :

- la saisie des métadonnées,
- le classement des documents par domaines scientifiques (ex : biologie, aquaculture, pêche, …),
- l'ajout de mots-clés, si utile,
- la remise en forme du texte intégral et sa conversion en PDF si nécessaire,
- le transfert du texte intégral vers le serveur d'Archimer.

a) Enregistrement des thèses, des actes de congrès ou des rapports internes

Pour les thèses, les actes de congrès ou les rapports internes, ce sont les auteurs eux-mêmes qui nous proposent l'enregistrement de leurs documents.

Pour diffuser un document de ce type via Archimer, les auteurs doivent donc nous fournir, par email, les données bibliographiques nécessaires au référencement de leur document. Ils doivent également envoyer le texte intégral de leur document, par email, CDROM ou au format Word ou au format PDF, en fonction de la taille du fichier.

Si les auteurs fournissent le texte intégral de leur document sous forme d'un ou de plusieurs fichiers Word, nous les convertissons et les fusionnons en un fichier PDF unique, à l'aide du logiciel Acrobat, avant transfert sur le serveur d'Archimer.

b) Enregistrement des publications récentes

Quelques auteurs nous signalent eux-mêmes les publications qu'ils souhaitent voir diffuser à partir d'Archimer. Dans ce cas, nous vérifions quelles sont les règles fixées par l'éditeur de la publication en matière d'auto-archivage. Si l'éditeur autorise cet auto-archivage, nous indiquons aux auteurs les éléments nécessaires à l'enregistrement de ces publications.

Toutefois, pour permettre l'enregistrement et la diffusion d'un plus grand nombre de publications, nous ne comptons pas uniquement sur des dépôts spontanés de la part des auteurs Ifremer, mais nous procédons nous-mêmes aux veille et collectes suivantes :

- Toutes les semaines, nous repérons les publications rédigées par le personnel de l'Ifremer. Toutes ces publications sont enregistrées dans la base « Bibliométrie » de l'Ifremer (voir chapitre suitant).

- Nous étudions ensuite la politique de chacun des éditeurs des publications Ifremer repérées, et ce à l'aide, notamment, du site WEB Sherpa/Romeo. Si la politique de l'éditeur n'est pas déclarée, ni sur le site Sherpa/Romeo, ni sur son propre site WEB, nous essayons systématiquement alors de contacter l'éditeur pour lui demander l'autorisation d'enregistrer ces articles dans Archimer.

- Si l'éditeur autorise l'auto-archivage de ses propres fichiers PDF (ex : EDP Sciences, The Company of Biologists…), nous déchargeons nous même le fichier PDF correspondant à l'article considéré à partir du site de l'éditeur et nous l'enregistrons dans Archimer. La majorité des données bibliographiques sont automatiquement transférées de la base « Bibliométrie » à la base Archimer. Les données bibliographiques manquantes sont copiées manuellement à partir du site de l'éditeur. Dans ce dernier cas de figure, l'enregistrement est donc effectué sans qu'il soit nécessaire de contacter les auteurs.

- Si l'éditeur autorise l'auto-archivage, mais limite cette dérogation à son droit de copyright au dernier « draft » de l'auteur (c'est à dire la version envoyée par l'auteur à l'éditeur : version qui contient toutes les corrections demandées par les pairs lors du processus de relecture, mais n'a pas été mise en page par l'éditeur), nous contactons les auteurs de la publication pour leur demander cette version. S'ils sont en mesure de nous la fournir, nous produisons nous mêmes, à partir des fichiers envoyés par les auteurs, le fichier PDF correspondant à nos critères de diffusion, et ce avant de l'enregistrer dans Archimer. Deux cas se présentent :

    - Soit l'auteur nous transmet sa publication sous la forme d'un ou de plusieurs fichiers Word (un pour le texte, un autre pour les tableaux et les figures par exemple), nous fusionnons ces fichiers avant de les convertir en un fichier PDF unique. Nous reconstruisons également la première page de la publication, non seulement pour uniformiser la présentation de nos publications, mais également pour répondre aux

règles fixées par les éditeurs (ajout de la citation complète et normalisée de la citation de la publication, ajout d'un lien vers le site de l'éditeur, ajout d'un texte explicatif spécifique à chaque éditeur…)

- Soit l'auteur nous transmet sa publication en format PDF, nous reconstruisons la première page avant de l'enregistrer dans Archimer.

2.3. Une archive liée aux autres systèmes documentaires de la bibliothèque La Pérouse

La figure n°1 présente l'organisation technique des principaux modules du système Archimer et leurs liens avec d'autres systèmes documentaires de la bibliothèque.

Archimer est en effet, par exemple, lié à la base Bibliométrie (voir fig. 1/5) de l'Ifremer. L'objectif de cette base est de répertorier les articles publiés, par le personnel de l'Ifremer, dans des revues à comité de lecture. Cette base de données a été développée, dans le contexte de la mise en place d'indicateurs nationaux d'évaluation de la production scientifique des organismes de recherche français. Elle est alimentée à partir de croisement de données exportées de la base de données Current Contents Connect® et de l'annuaire LDAP de l'Ifremer (voir fig. 1/6 et 1/7).

Pour simplifier les contacts avec les auteurs de ces publications, le personnel de la bibliothèque dispose d'Archimail (voir fig. 1/3) qui utilise les données de cette base Bibliométrie. Cet outil permet la génération de messages pré-rédigés et personnalisés en fonction de la publication à traiter. Quand, dans la base Bibliométrie (voir fig. 1/5), le personnel de la bibliothèque détecte un article publié dans une revue, dont l'éditeur autorise l'auto-archivage, il suffit en effet de copier l'identifiant de l'article et de la copier dans le formulaire d'accueil d'ArchiMail. A l'aide de cet identifiant, Archimail récupère, dans la base Bibliométrie les données nécessaires à la composition du message, et, notamment, le titre de la publication et l'adresse email de tous les auteurs Ifremer repérés dans la publication considérée. A l'aide des ces informations, Archimail compose ensuite un message, qu'il est possible de personnaliser, avant de l'envoyer automatiquement à tous les auteurs repérés dans cet article.

Pour enregistrer un nouveau document dans Archimer, le personnel de la bibliothèque se connecte à un site WEB (voir fig. 1/4) accessible à partir de l'Intranet de l'Ifremer. Ce site WEB propose un ensemble de formulaires WEB spécifiques aux types de documents à enregistrer (Thèses, rapports internes, publications…).

Ces formulaires permettent l'enregistrement des données bibliographiques du document (titre, résumé, auteur…) qui seront sauvegardées dans une base de données (voir fig. 1/1). Ces formulaires permettent également l'enregistrement du texte intégral, sous forme d'un fichier PDF, stocké sur un espace disque du serveur Internet de l'Ifremer (voir fig. 1/2).

Pour enregistrer une publication déjà référencée dans la base Bibliométrie (voir fig. 1/5), le personnel de la bibliothèque peut saisir l'identifiant associé à ce document dans cette base de données. Cette option permet de transférer automatiquement les données bibliographiques disponibles dans cette base de données vers les formulaires de saisie. Dans ce cas, pour le personnel considéré, il ne reste plus, pour compléter l'enregistrement, qu'à saisir les informations manquantes (DOI, copyright, texte intégral).

Dans le cas de l'enregistrement d'une publication, pour obtenir automatiquement les données relatives à la revue dans laquelle l'article est publié, le personnel de la bibliothèque peut également faire une recherche dans la base « revues électroniques » (voir fig. 1/8). Cette base de données contient la liste cumulative de tous les titres auxquels Ifremer a souscrit un abonnement. Le personnel peut ainsi saisir quelques mots du titre (ex : aqua* liv*), pour retrouver le titre correspondant, et transférer l'ensemble de ces données dans les formulaires de saisie (titre complet de la revue, URL de la revue, nom de l'éditeur, URL de l'éditeur).

Les utilisateurs extérieurs peuvent consulter les documents disponibles via le site WEB d'Archimer (voir fig. 1/11). Dans ce cas, les notices des documents enregistrés sont construites dynamiquement à l'aide de pages JSP en fonction des requêtes exécutées par les utilisateurs. Ces notices proposent des liens vers le texte intégral des documents (voir fig. 1/2).

Toutes les nuits, un programme JAVA (voir fig. 1/9) construit un fichier HTML statique pour chaque nouveau document enregistré (voir fig. 1/10). Ce fichier statique propose la notice du document ainsi qu'un lien vers le texte intégral (voir fig. 1/2). Ces fichiers statiques sont construits à l'intention des robots

des moteurs de recherches Internet (ex : Google, MSN). De cette façon, ces notices et le texte intégral des documents sont directement accessibles à partir de ces moteurs de recherche.

Archimer est également compatible OAI-PMH. Les moissonneurs, et notamment Avano, décrit à la suite de ce document (voir fig. 1/13), peuvent donc récolter les données bibliographiques des données enregistrées dans Archimer en interrogeant son serveur OAI-PMH (voir fig. 18.10). Ce moissonneur pourra ainsi alimenter sa propre base de données bibliographique (voir fig. 1/12) à partir des références récoltées dans plusieurs archives. Ces moissonneurs, offriront ainsi, à partir de leur propre interface de consultation, un accès aux notices statiques d'Archimer (voir fig. 1/10).

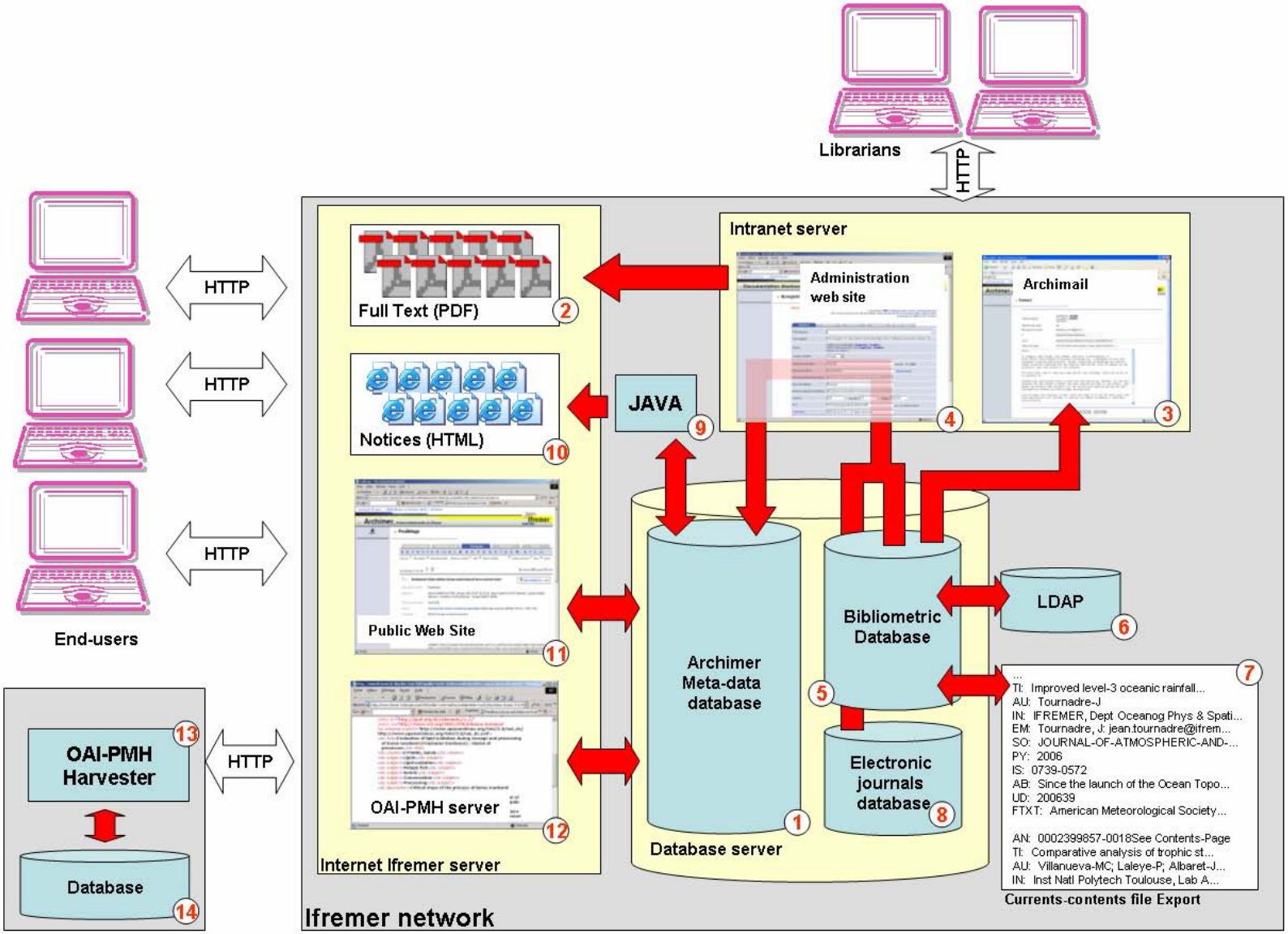

Figure n°1 : architecture du système Archimer

2.4. Résultats d'une année de collecte des publications à l'Ifremer

Une année après son ouverture Archimer propose plus de 1400 documents dont déjà plus de 70% des publications rédigées ou co-rédigées par l'Ifremer depuis août 2005.

En effet, du 1° août 2005 au 15 août 2006, 116 publications avec un premier auteur Ifremer ont été repérées dans la base de données Current Contents Connect®. 82 de ces 116 publications sont déjà enregistrées, soit près de 70%. Ces 116 publications peuvent être réparties de la façon suivante :

- 10 articles ont été publiés chez des éditeurs qui interdisaient l'enregistrement de leurs publications dans une archive institutionnelle (ex : American Meteorological Society, ASLO...),

- 16 articles ont été publiés chez des éditeurs qui autorisaient l'auto-archivage de leurs propres fichiers PDF. 8 de ces 16 articles sont toujours sous embargo. Ils sont enregistrés mais ne seront visibles que dans quelques mois, ce qui devrait porter rapidement le pourcentage de publications en libre accès à plus de **77%**,

- 90 articles ont été publiés chez des éditeurs qui limitaient le droit d'auto-archivage au dernier draft de la publication. Les drafts de 74 de ces 90 articles ont pu être collectés et enregistrés.

D'une manière plus large, pendant la même période, 257 publications avec un ou plusieurs auteurs Ifremer, quelque soit la ou leur position dans la liste des auteurs, ont été repérées dans la base de données Current Contents Connect®. Plus de 60% de ces 257 publications sont déjà en accès libre via Archimer.

2.5. Perspectives d'évolutions

2.5.1. Collecte des « versions auteurs » de publications dès l'acceptation par l'éditeur

Pour l'instant, peu d'auteurs déposent leurs documents spontanément dans Archimer. Nous avons obtenu la majorité des documents actuellement enregistrés dans Archimer en contactant nommément leurs auteurs. Cependant, cette méthode présente plusieurs limites :

- Lorsque nous cherchons à contacter les auteurs d'une publication, ils ont parfois déjà quitté l'Ifremer. Cela s'explique par le fait qu'il peut s'écouler plus d'une année entre la soumission d'un article à une revue et son apparition dans les Current Contents. Quand un thésard publie un article pour présenter ses travaux à la fin de sa thèse, il a donc souvent quitté l'Ifremer au moment où son article est publié et qu'il apparaît dans les Current Contents,

- Si un éditeur autorise uniquement la diffusion du dernier « Draft » d'une publication, au moment où nous contactons les auteurs pour obtenir cette version, il est parfois trop tard, car ces fichiers ont été perdus ou supprimés.

Nous avons donc commencé à mettre en place une collecte systématique des « versions auteurs » dès leur acceptation par une revue. En effet, quand les auteurs nous déposent leurs fichiers dès l'acceptation de leur publication par une revue, nous pouvons, non seulement améliorer le pourcentage de collecte des publications Ifremer, mais surtout, sous réserve d'une compatibilité avec les règles de copyright fixées pas l'éditeur, diffuser les publications « In-Press ». Nous pouvons ainsi participer à l'accélération du processus de diffusion des résultats des recherches de l'Ifremer, en diffusant les publications plusieurs mois avant qu'elles n'apparaissent sur le site Internet de l'éditeur.

2.5.2. Elargissement du système à d'autres types de documents

Actuellement, Archimer permet l'enregistrement et la diffusion de thèses, de post-publications, de publications In-press, de rapports internes, de rapports d'activités et d'actes de congrès. Nous envisageons l'intégration d'autres types de documents :

- Brevets,
- Posters,
- HDR
- …

3. Avano, un moissonneur OAI pour les sciences marines et aquatiques

3.1. Contexte

Une année après le lancement d'Archimer, la Bibliothèque La Pérouse lance Avano, un moissonneur OAI spécialisé en sciences marines et aquatiques. Ce développement a pour objectifs:

- De continuer à afficher le soutient d'Ifremer au mouvement Open Access
- D'offrir une meilleure visibilité aux documents déposés dans Archimer, et ce, en les agrégeant avec les travaux déposés dans plusieurs autres archives, afin de constituer une base d'envergure internationale.
- D'offrir à la communauté scientifique des sciences de la mer un nouvel outil centralisé pour découvrir des données aujourd'hui disséminées sur un ensemble de serveurs.

3.2. Principe de fonctionnement

Avano est un moissonneur OAI pour les sciences marines et aquatiques. Il collecte donc les données bibliographiques des ressources électroniques (documentation, images, données brutes...) disponibles dans un ensemble d'Archives Ouvertes via le protocole OAI-PMH pour les agréger dans une base de données centralisée (voir fig. 2/3). Il offre ainsi à ses utilisateurs une interface WEB (voir fig. 2/3) qui permet de repérer de façon centralisée des ressources disséminées dans un ensemble de serveur.

Avano moissonne ainsi plusieurs archives d'instituts de recherches en sciences marines. Toutes les ressources stockées dans ces archives spécialisées en sciences marines sont systématiquement et automatiquement référencées dans Avano. Fin septembre 2006, Avano moissonne ainsi les 6 archives spécialisées en sciences marines suivantes :

Archive	Nb. Doc. Disponibles	Description
ArchiMer, Institutional Archive of Ifremer (French Research Institute for Exploitation of the Sea)	1446	Archimer, is the institutional repository of Ifremer (French Research Institute for Exploitation of the Sea). It provides freely available scientific or technical documents online (publications, theses, conference proceedings, etc) in all fields related to oceans (oceanography, aquaculture, fisheries, etc).
DRS at National Institute Of Oceanography	418	The National Institute of Oceanography (NIO) in India hosts the Digital Repository Service (DRS) which collects, preserves and disseminates institutional publications (journal articles, conference proceedings, technical reports, theses, dissertations, etc.).
Marine & Ocean Science ePrints Archive @ Plymouth	1520	Marine & Ocean Sciences ePrints @ Plymouth is a digital archive providing access to papers produced by the staff of the Marine Biological Association of the United Kingdom, Plymouth Marine Laboratory and the Sir Alister Hardy Foundation for Ocean Science. Marine & Ocean Sciences ePrints @ Plymouth is also an historical archive containing digital copies of early papers from the Journal of the Marine Biological Association of the United Kingdom.
OdinPubAfrica	1112	Research & Publications in Marine Science in Africa in digital form, including preprints, published articles, technical reports, working papers and more
Repository@NOAA	34	Repository@NOAA (the National Oceanic and Atmospheric Administration) is a searchable database of full-text, online NOAA documents from several selected NOAA programs. The purpose of this project is to establish the feasibility and importance of archiving for the long-term full-text NOAA documents in a secure, accessible, and authenticated NOAA electronic repository. The NOAA IR Pilot Project is a collaboration between the NOAA Libraries and Information Network, the NOAA Central Library, and the Digital Commons Institutional Repository platform developed by Berkeley

		Electronic Press.
WHOAS at MBLWHOI Library	1190	Repository of the Woods Hole Scientific Community, covering the subjects of ocean physics and engineering, oceanography and marine biology

Avano interroge également un ensemble d'archives ouvertes non spécialisé en sciences marines dans lesquelles sont stockées, parmi d'autres, un ensemble de ressources liés aux sciences marines et aquatiques. C'est le cas, par exemple, du serveur ArXiv spécialisé en sciences physiques et mathématiques qui contient un ensemble de publications liées à l'océanographie.

Quelques une de ces archives offrent la possibilité d'isoler les documents liés aux domaines qui nous intéressent à partir de subset. Dans ce cas, il est possible d'isoler automatiquement les ressources liées aux sciences marines ou aquatiques et le rendre automatiquement visible au public d'Avano.

Pour traiter les archives qui n'offrent pas une classification parfaitement adaptée aux domaines qui nous intéressent (voir fig. 2/5), Avano décharge (voir fig. 2/6) la totalité de leurs notices dans une base de données temporaire (voir fig. 2/8).

Ces données sont indexées et un système automatique (voir fig. 2/9) isole les notices qui contiennent un ou plusieurs termes liés aux sciences marines ou aquatiques (voir fig. 2/10).

Les notices que le système repère par ce système (voir fig. 2/11) de mots-clés sont ensuite validées, manuellement, par le personnel de la bibliothèque (voir fig. 2/12) avant d'être visibles via Avano. Pour valider ces notices, le personnel de la bibliothèque dispose d'un site WEB (voir fig. 3). Les mots-clés repérés dans les notices sont surlignés. Ce système permet au personnel de la bibliothèque de rejeter les fiches quand ces mots clés sont employés dans un domaine autre ceux qui nous intéressent (par exemple quand le mot *Fish* est employé pour *Fluorescence in situ hybridization*).

Ce système de recherche par mot-clés nous a permis de publier, fin septembre 2006, plus de 25 000 notices isolés dans un ensemble de plus de 1.5 million de notices déchargé à partir de 35 archives non spécialisé en sciences marines.

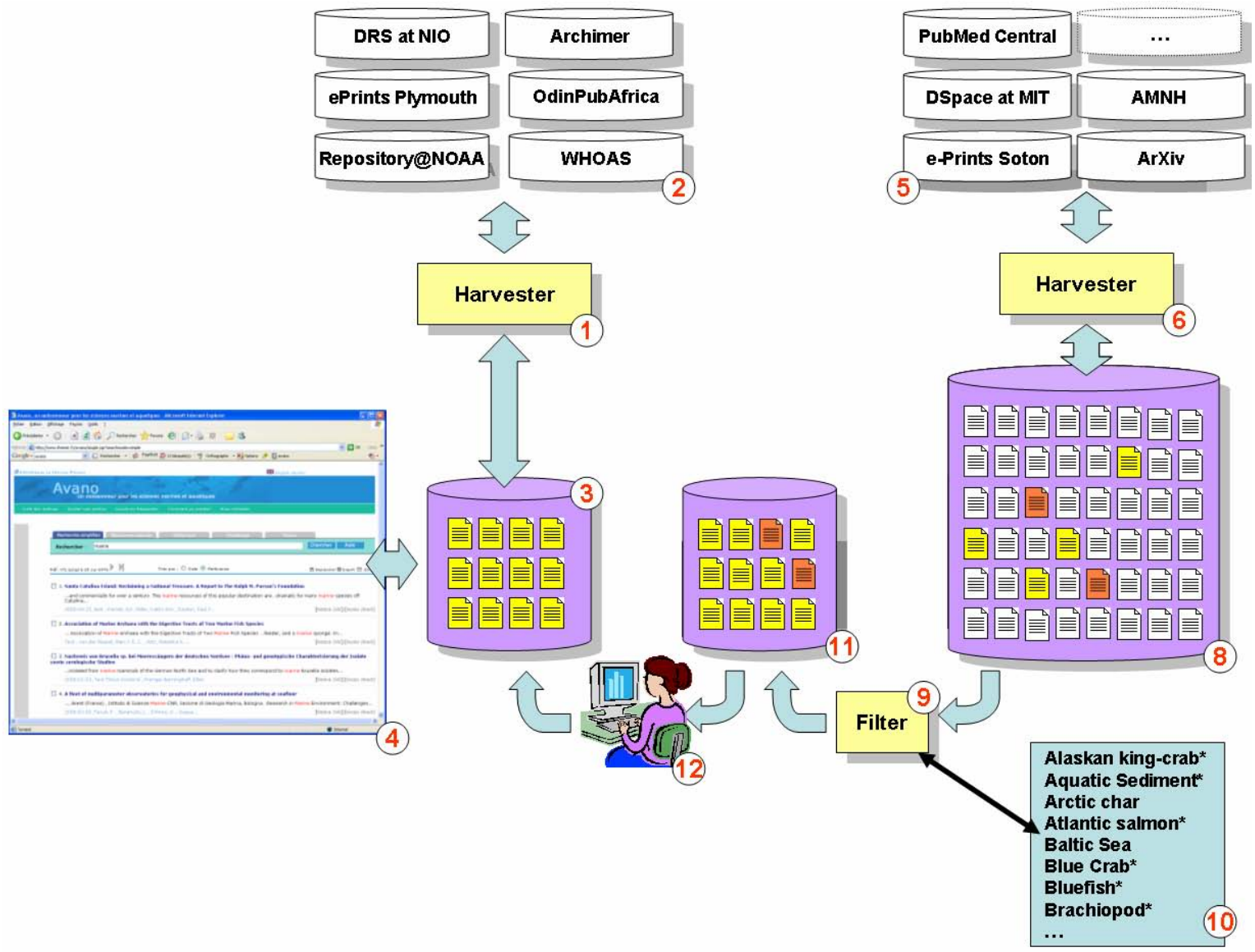

Figure n°2 : Principe de fonctionnement d'Avano

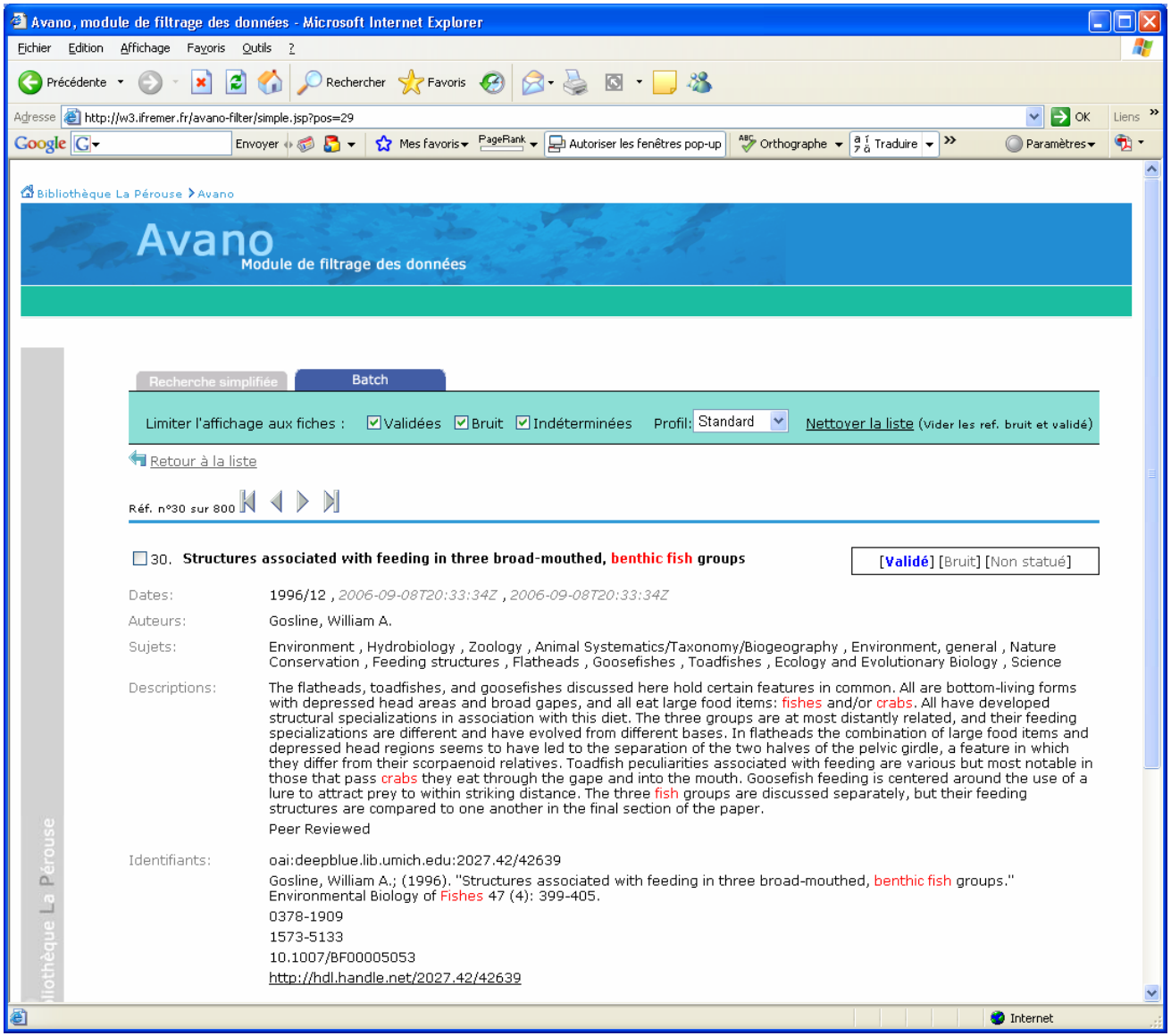

Figure n°3 : Module de filtrage des données d'Avano

3.3. Difficultés rencontrées lors la mise en œuvre d'Avano

Les difficultés que nous avons rencontrées dans la mise en œuvre de cette archive sont principalement liées aux limitations du protocole OAI-PMH :

- **Repérage des fiches correspondantes à une thématique dans une archive** : C'est la principale difficulté à la quelle nous ayons été confrontés. Il n'existe presque jamais de subset parfaitement adapté aux domaines que nous souhaitions isoler dans les archives non spécialisées en sciences marines. Cette limitation nous a conduit à développer le système de repérage par mots-clés décrit dans le chapitre précédent.

- **Gestion des fiches supprimées** : Certaines archives ne gardent pas des traces des fiches qu'elles suppriment de leur base. Ces archives sont donc incapables d'indiquer aux moissonneurs quelles fiches ont été supprimées. Dans ce cas, les moissonneurs, dont Avano, peuvent proposer des fiches qui pointent vers des ressources qui n'existent plus. Pour contourner ce problème, Avano devra re-moissonner complètement ces archives à intervalle régulier pour détecter d'éventuelles suppressions.

- **Gestion des doubles** : Plusieurs organismes de recherches ou universités peuvent enregistrer la même ressource électronique dans leur propre archive institutionnelles. Si Avano moissonne ces archives, il obtiendra des fiches descriptives du même objet stocké à plusieurs endroits. Ce cas peut par exemple survenir quand une publication est rédigée en collaboration avec plusieurs institutions. Dans ce cas, cette publication peut être archivée sur les différents serveurs de ces institutions. Vu le faible taux d'auto-archivage actuel (environ 15%), la probabilité d'afficher des doubles dans la liste de résultats reste encore faible, mais ce problème devrait prendre plus d'importance dans les années à venir.

- **Détermination des dates de publications et/ou des types de ressources**: Pour respecter le protocole OAI-PMH les archives sont tenues d'exposer leurs données dans la DTD Dublin-Core non qualifiée. Dans cette DTD tous les champs sont optionnels. Ce caractère facultatif des informations pose plusieurs problèmes et notamment pour les champs « date » et « type ». En effet, quand une fiche ne contient pas de date de publication, elle se retrouve systématiquement en fin de liste quand l'utilisateur demande à trier sa liste de résultats par date. De même, quand un utilisateur demande à limiter une recherche à une plage de dates spécifiques, ces fiches sont exclues de la recherche même si elles correspondent à la recherche spécifiée.

- **Normalisation du champ type** : Même si la DTD Dublin Core suggère de stoker l'information « type de document » à l'aide de chaînes de texte normalisées, peu d'archives respectent ce conseil et présentent cette information sous forme de texte libre (ex : « publication », « artjournal », « text », « article » sont utilisés pour décrire un article). Dans Avano, nous avons souhaité proposer aux utilisateurs de limiter leurs recherches à un ou plusieurs types de ressources (documentation, image, jeux de données, vidéo, audio). Pour permettre la mise en place de ce filtre nous avons donc du mettre en place un système de normalisation de cette données basé sur la reconnaissance de mots-clés dans cette chaîne de caractère. Cette normalisation n'est donc pas parfaite, et notre système de filtre peut donc exclure par erreur des ressources de la liste de résultats quand un utilisateur limite une recherche à un ou plusieurs types de données spécifiques.

3.4. Perspectives d'évolutions

Dans les prochains mois, Avano devraient moissonner d'Avantage d'Archives Ouvertes, dont, nous l'espérons, de nouvelles archives développées par les membres de Iamslic, et proposer ainsi à ses utilisateurs un plus grand nombre de notices.

De plus nous pourrions envisager de moissonner également le catalogue d'éditeurs privés. En effet, dès à présent, deux éditeurs (« HighWire Press » et « The University of Chicago Press Journals Division ») exposent déjà leurs publications en OAI-PMH. Si d'autres éditeurs adoptent également ce protocole OAI-PMH, nous pourrions envisager d'intégrer une sélection de leurs notices, dont le texte intégral restera accessible sous condition d'abonnement, en permettant aux utilisateurs de les filtrer, et de les agréger avec les travaux accessibles gratuitement à partir du réseau d'Archives Ouvertes.

Avano permettrait ainsi d'obtenir rapidement une vue plus complète de la recherche internationale dans les domaines des sciences marines et aquatiques.

3.5. Proposition de collaboration dans le cadre de IAMSLIC

Le lancement d'Avano nous a donné la satisfaction de constater l'intérêt de plusieurs de nos collègues de IAMSLIC, dont les initiateurs du projet « Aquatic Commons », pour ce système. Nous espérons d'ailleurs qu'Avano pourra s'intégrer dans ce projet. Dans cette perspective, nous souhaitons proposer aux membres projet « Aquatic Commons », voire à d'autres collègues de IAMSLIC, de s'associer à la mise en œuvre de ce système et notamment à sélection des notices issues des archives non-spécialiées en sciences marines et aquatiques.

Références bibliographiques

Documentation :

Le protocole OAI et ses usages en bibliothèque
François NAWROCKI - Ministère de la culture et de la communication
28 janvier 2005 - http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm

Archimer, ou la mise en place d'une Archive Institutionnelle à l'Ifremer
Fred Merceur - 23 novembre 2005
http://www.ifremer.fr/docelec/doc/2005/rapport-657.pdf

Sites Web :

Site de l'Open Archive Initiative
http://www.openarchives.org/

Oaister
http://oaister.umdl.umich.edu/o/oaister/

OAI explorer
http://re.cs.uct.ac.za/