33nd IAMSLIC Annual Conference
Changes on the Horizon
October 7-11, 2007
Sarasota, Florida, USA

Fred Merceur
frederic.merceur@ifremer.fr
Ifremer – Bibliothèque La Pérouse
BP 70, 29280 Plouzané, France
V1.0

Avano, bilan d'une année de gestion d'un moissonneur OAI-PMH thématique

Résumé: En Septembre 2006, la Bibliothèque La Pérouse lançait Avano, un moissonneur OAI pour les sciences marines et aquatiques. Avano propose aujourd'hui un accès centralisé à plus de 100.000 références, dont une grande majorité de documents accessibles gratuitement en texte intégral, moissonnés à partir de plus de 150 Archives Ouvertes.

Près d'une année après son ouverture, ce document a pour objectif de rappeler le principe de fonctionnement de ce moissonneur thématique. Ce document est également l'occasion de faire le bilan des principales difficultés rencontrées lors de cette première année de gestion d'Avano, et, en conclusion, d'envisager quelques solutions pour permettre d'améliorer la qualité du service offert par Avano à ses utilisateurs.

Mots-clés: Open Access, Protocole OAI-PMH, Libre accès, Archives Ouvertes, Archives Institutionnelles, Moissonneur.

Table des matières

1. Introduction 2

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

2.1. Informations générales 3
2.2. Principe de fonctionnement 3
2.3. Bilan d'une année de fonctionnement 6
2.3.1. Une année de moissonnage 6
2.3.2. Statistiques de consultations 8

3. Difficultés liées à l'implémentation de certaines archives et limites du protocole OAI-PMH 9

3.1. Stabilité des archives 9
3.2. Erreurs de structures des flux XML et d'encodage des caractères UTF8 9
3.3. Moissonnage des grosses archives 9
3.4. Gestion des doubles 10
3.5. Gestion des notices supprimées 10
3.6. Gestion du champ Type 10
3.7. Gestion du champ Date de publication 11
3.8. Notices pauvres 12
3.9. Mélange données brutes / documentation 12
3.10. Notices sans accès gratuit à l'objet numérique 14
3.11. Moissonnage thématique 14

4. Conclusion 14

5. Références 16

1. Introduction

Depuis le début des années 90 des communautés scientifiques ont créé des serveurs de pré-publications pour offrir un accès gratuit et immédiat à leurs travaux (ex : ArXiv en physique).

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 (ou moissonneur), dont Avano, 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.

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

2.1. Informations générales

Avano est un moissonneur OAI pour les sciences marines et aquatiques, accessible à partir de l'adresse http://www.ifremer.fr/avano/. Avano propose, en septembre 2007, un accès centralisé à plus de 100.000 références de ressources électroniques, dont une grande majorité de documents accessibles gratuitement en texte intégral.

Avano propose un accès à des ressources liées aux sciences marines (pêche, aquaculture, biologie marine, géologie marine, économie marine, océanographie, écologie marine…) ainsi qu'à des ressources liées aux ressources en eau douce (gestions des lacs et des rivières, restauration des zones humides, traitement des eaux usées…).

Avano est en partie fondé sur la version JAVA du système Open Archives Initiative Metadata Harvesting Project développée par l'Université de l'Illinois. Le système de filtre présenté dans le paragraphe suivant ainsi que le site public de consultation d'Avano ont été développés par l'Ifremer à l'aide des technologies JSP, JAVA et Oracle.

Avano moissonne aujourd'hui plus de 150 archives ainsi que 4 éditeurs commerciaux. Avano moissonne non seulement des archives spécialisées en sciences aquatiques mais également un ensemble d'archives généralistes. Pour isoler les notices liées au milieu aquatique disponibles dans ces archives généralistes, Avano utilise un système de recherche de mots-clés décrit dans le paragraphe suivant.

Dans la mesure du possible nous n'enregistrons dans Avano que les archives qui proposent une grande majorité de notices avec un lien, gratuit ou payant, vers l'objet numérique. Nous essayons donc d'éviter d'enregistrer les archives proposant une majorité de notices vides, sans aucun lien vers la ressource. Cette mesure ne s'applique évidement pas si l'archive propose un *Set* permettant le moissonnage des seules notices proposant un lien vers la ressource.

Dans la mesure du possible également, nous n'enregistrons pas les archives proposant des notices qui pointent vers des ressources situées à l'extérieur de leur serveur et notamment les archives qui référencent des ressources collectées sur le WEB.

2.2. Principe de fonctionnement

Avano est un moissonneur OAI. Il collecte donc les données bibliographiques des ressources électroniques (documentation, images, données brutes...) disponibles dans un ensemble de réservoirs via le protocole OAI-PMH pour les agréger dans une base de données centralisée. Il offre ainsi à ses utilisateurs une interface WEB qui permet de repérer de façon centralisée des ressources disséminées dans un ensemble de serveurs.

Avano moissonne plusieurs archives d'instituts de recherches en sciences aquatiques. Toutes les ressources stockées dans ces archives spécialisées sont systématiquement et automatiquement référencées dans Avano. En septembre 2007, Avano moissonne ainsi 9 archives spécialisées en sciences aquatiques (Fig. 1). Près de 19.000 notices disponibles dans Avano sont directement issues de ces 9 archives.

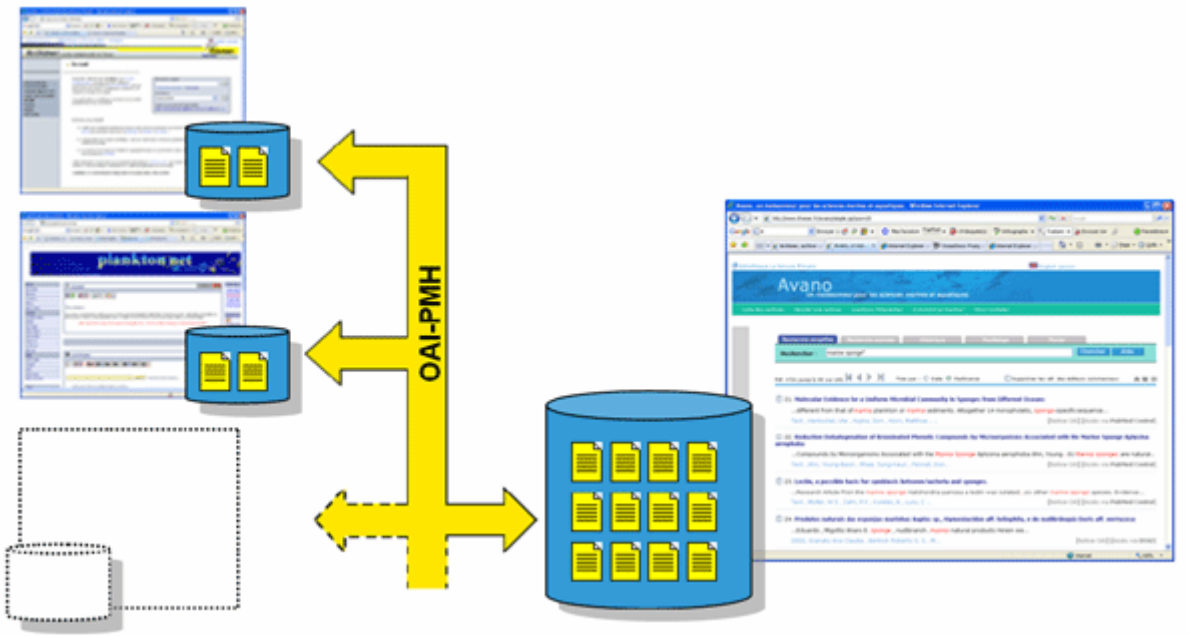

Fig. 1. Avano moissonne, en septembre 2007, 9 archives spécialisées en sciences marines. Les notices exposées par ces 9 archives sont systématiquement enregistrées dans Avano.

Avano interroge également un ensemble d'archives ouvertes non spécialisées en sciences aquatiques dans lesquelles sont stockées, parmi d'autres, un ensemble de ressources intéressantes. C'est le cas, par exemple, du serveur PubMed Central spécialisé en sciences biomédicales et en science du vivant. PubMed Central propose ainsi plus de 1.000.000 de documents dont 15.000 correspondent aux domaines de recherche d'Avano.

Si le moissonnage thématique d'une archive est supposé possible via le mécanisme des *Set* du protocole OAI-PMH, dans la réalité nous n'avons bien sûr jamais trouvé de *Set* «sciences marines et aquatiques» dans aucune des archives que nous avons moissonnées. Pour filtrer ces archives nous avons donc développé un système de recherche de termes et d'expressions relatifs aux sciences aquatiques.

Pour traiter les archives qui n'offrent pas une classification parfaitement adaptée aux domaines qui nous intéressent (Fig. 2.1), Avano décharge la totalité de leurs notices dans une base de données temporaires (Fig. 2.2).

Ces données sont indexées et un premier système automatique (Fig. 2.3) recherche près de 30.000 noms scientifiques d'espèces aquatiques dans l'intégralité de la notice. Si une notice contient, par exemple, la chaîne de caractère *Crassostrea gigas* (nom scientifique d'une espèce d'huître), nous considérons qu'il a peu de chance que ce nom soit utilisé dans un contexte autre que ceux qui nous intéressent. La notice sera donc automatiquement visible dans Avano (Fig. 2.4). Cette liste de 30.000 noms d'espèces a été compilée à partir de listes proposées par le projet FishBase, par la FAO ainsi que par la NODC.

Et je profite de la rédaction de ce papier pour lancer un appel aux détenteurs de listes de noms scientifiques d'espèces aquatiques et notamment d'espèces d'algues, de mousses, de plantes, de mollusques, de gastéropodes, d'insectes, d'oiseaux et de mammifères. Si vous détenez de telles listes et si elles ne sont pas mélangées avec des espèces non aquatiques, pouvez vous me contacter ? Ces listes nous seraient très utiles pour continuer à automatiser notre processus de filtres des notices.

Ce premier système recherche également plus de 1.000 noms de journaux ou de bulletins spécialisés en sciences aquatiques dans le champ *Source* de la notice. Si Avano détecte un de ces journaux dans le champ *Source* d'une notice, elle est automatiquement enregistrée dans Avano.

Avano recherche également un ensemble de termes et d'expressions plus généraux liés au milieu aquatique (Fig. 5). Avano recherche par exemple les mots *Poisson*, *Marine*, *Pêche*, *traitement de l'eau*… Les notices que le système repère par ce système (Fig. 2.5) de mots-clés sont ensuite validées,

manuellement, par le personnel de la bibliothèque (voir Fig. 2.6) avant d'être visibles via Avano. Un site WEB permet de valider ces notices (voir Fig. 3). Les mots-clés repérés dans les notices sont surlignés. Ce système permet 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 filtre basé sur la recherche de mots-clés nous a permis de récolter, fin septembre 2007, plus de 88.000 références isolées dans un ensemble de plus de 4.5 millions de notices déchargées à partir de 146 archives non spécialisées en sciences aquatiques.

Cette méthode est bien sur loin d'être idéale :

- Cette méthode repose en partie sur un tri manuel des notices qui prend du temps (quelques minutes par jour pour filtrer les nouvelles fiches des 150 archives déjà enregistrées sans compter le temps pour traiter le *back-log* quand nous enregistrons de nouvelles archives),
- Comme nous ne consacrons généralement pas plus de 2 ou 3 secondes pour valider ou non une fiche, nous enregistrons sans doute, par erreur, un faible pourcentage (1 à 2% ?) de fiches qui ne correspondent pas aux domaines couverts par Avano,
- Nous passons sans doute également à coté d'un faible pourcentage de notices (1 à 2% ?), et particulièrement quand les notices sont pauvres (quand elles n'offrent ni mots-clés ni résumé) ou quand elles n'offrent que du texte dans une langue nationale.

C'est toutefois la seule méthode (que nous ayons imaginée qui nous permette de récupérer près de 80% des notices aujourd'hui disponibles dans Avano.

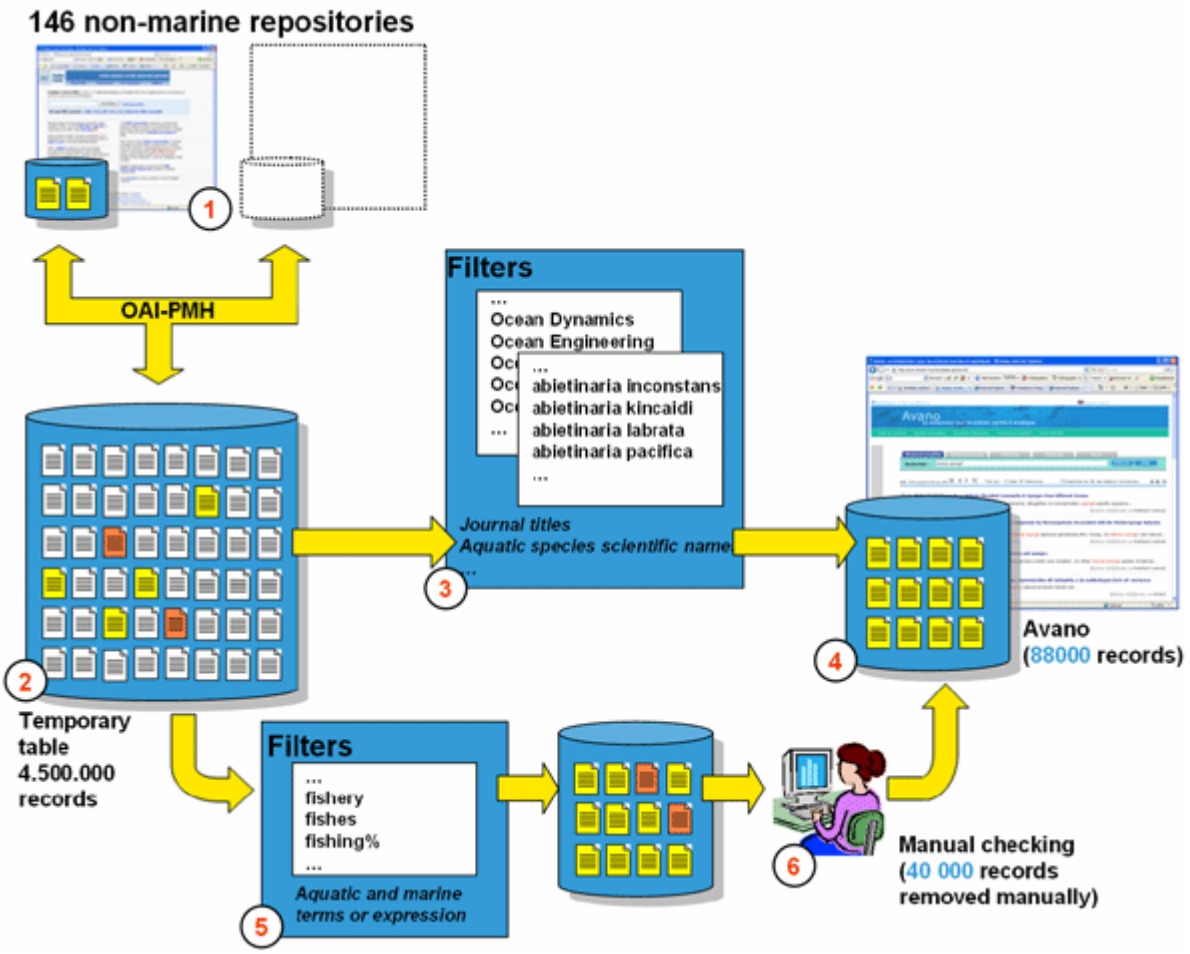

Fig. 2. Avano moissonne un ensemble d'archives généralistes. Les notices liées aux sciences aquatiques disponibles dans ces archives sont isolées à l'aide d'un système de recherche de mots-clés.

Fig. 3. Module de filtrage manuel des notices issues d'archives généralistes dans lesquelles ont été repéré un ou plusieurs termes relatifs aux sciences marines et aquatiques.

2.3. Bilan d'une année de fonctionnement

2.3.1. Une année de moissonnage

En septembre 2007, un an après son lancement, Avano propose un accès à plus de 107.000 ressources issues de plus de 150 Archives et de 4 éditeurs commerciaux. La figure 4 présente l'évolution du nombre de notices disponibles. La plus grande partie de cette progression correspond au chargement d'archives déjà existantes et donc au traitement du *BackLog*. L'année prochaine nous ne pouvons espérer une progression aussi importante puisque nous avons à présent enregistré la majorité des archives existantes répondant aux critères que nous nous sommes définis. La figure 5 est donc plus intéressante puisqu'elle présente le nombre de documents disponibles dans Avano par année de publication. Cette courbe prouverait un réel démarrage du mouvement Open Access puisque le nombre de documents récents accessible via le protocole OAI-PMH semble de plus en plus important chaque année avec une forte progression en 2006.

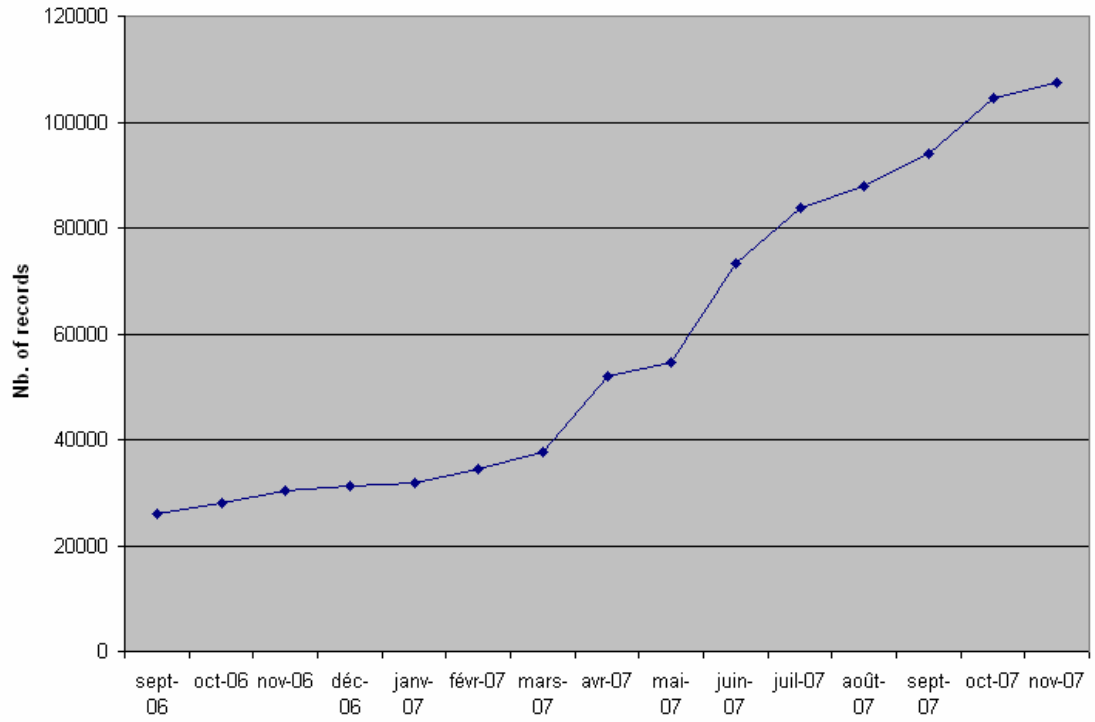

Fig. 4. Augmentation du nombre de notices disponibles à partir d'Avano depuis son lancement.

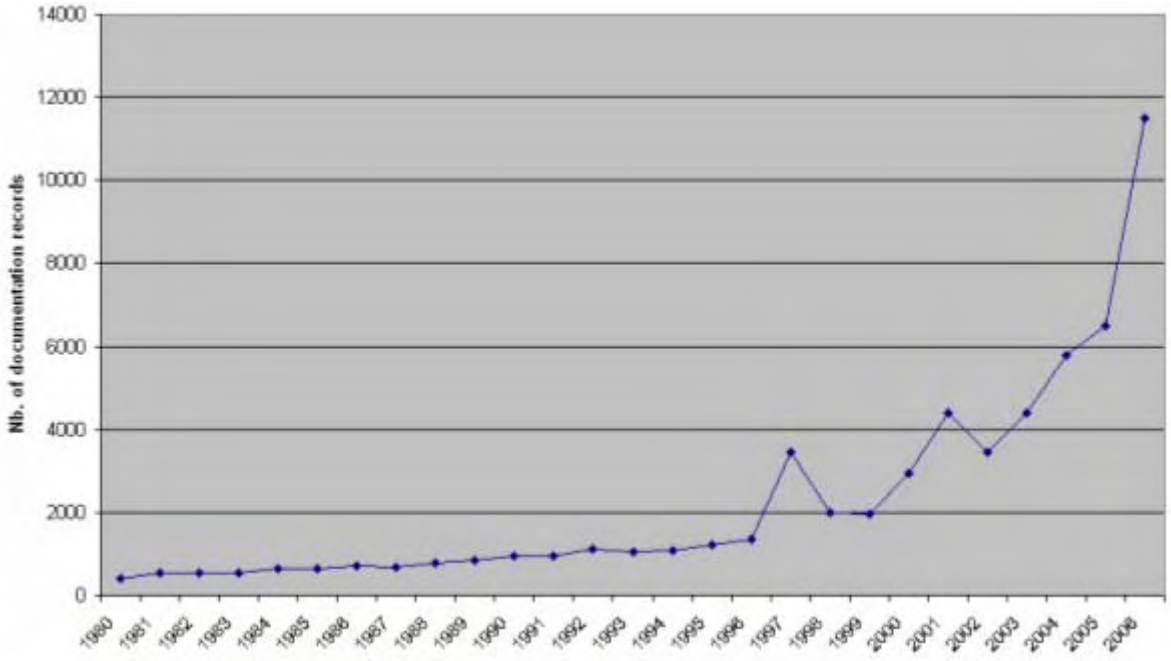

Fig. 5. Année de publication des documents accessibles à partir d'Avano. Seules les notices clairement associées à des documents sont pris en compte dans ce graphique (les notices sans indication de type, les images, les fichiers de données, par exemple, ne sont pas pris en compte dans ce graphique).

2.3.2. Statistiques de consultations

Si elles restent encore faibles, le nombre de connexions augmente pour l'instant régulièrement (Fig. 6). La majorité des connexions proviennent de la France (pour laquelle nous disposons de canaux de promotion) et des USA (Fig. 7).

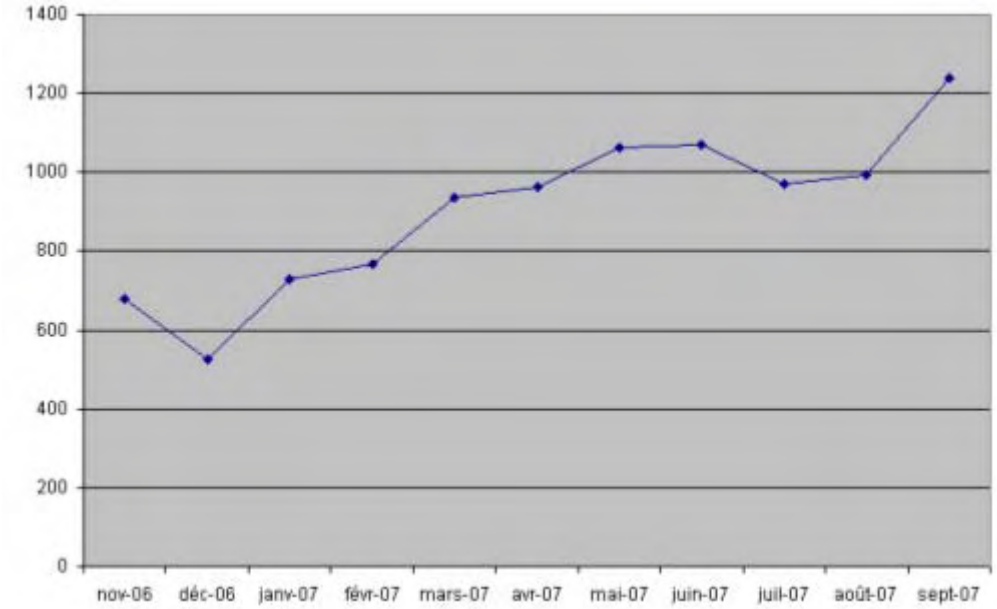

Fig. 6. Progression du nombre de connexions à Avano depuis son lancement

Fig. 7. Synthèse géographique, générée par Google analytics, des connexions à Avano pour le mois de Septembre 2007. La taille et la densité des points affichés sont proportionnelles aux nombres de connexions sur la période donnée.

3. Difficultés liées à l'implémentation de certaines archives et limites du protocole OAI-PMH

Depuis le lancement d'Avano, nous sommes confrontés à de nombreuses difficultés techniques qui rendent sa gestion plus difficile que prévue. Loin du système complètement automatisé et autonome que nous avions imaginé avant son lancement, l'administration d'Avano nécessite de nombreuses interventions manuelles. Les difficultés que nous avons rencontrées dans la mise en œuvre de ce moissonneur sont en partie liées aux limitations et à la trop grande permissivité du protocole OAI-PMH et à des problèmes d'implémentation de certaines archives.

De plus, certaines archives présentent des données de mauvaises qualités (ex : notice sans date de publication..). Et ces données nuisent malheureusement à la qualité globale du service offert par Avano à ses utilisateurs.

Ce paragraphe liste donc les principales difficultés rencontrées lors de cette première année de fonctionnement d'Avano.

3.1. Stabilité des archives

Certaines archives présentent aujourd'hui des problèmes de stabilité. Ces problèmes imposent un travail d'administration important pour permettre le moissonnage de ces archives. La liste ci-dessous correspond par exemple à une sélection de problèmes auxquels nous sommes confrontés régulièrement depuis le lancement d'Avano :

- Certains serveurs sont régulièrement inaccessibles ou retournent des erreurs non documentées,
- Certains moissonnages sont interrompus par des erreurs de *Time-Out* http, par des erreurs non documentées ou sans aucun message d'erreur,
- Certaines archives ne supportent le protocole OAI-PMH que de manière partielle. Certaines archives ne supportent, par exemple, que la méthode *GetRecords*. D'autres ne supportent que la méthode *ListIdentifiers+GetRecord*. D'autres encore ne retournent pas le même nombre de documents en fonction de la méthode sélectionnée!
- Les adresses de certains serveurs sont parfois modifiées sans que la communauté en soit avertie
- …

3.2. Erreurs de structures des flux XML et d'encodage des caractères UTF8

Certaines archives transmettent également des notices dans des flux XML non conformes à la DTD spécifiée. D'autres retournent des notices comportant des caractères UTF-8 non conformes. Ces erreurs sont problématiques pour certains moissonneurs, et notamment pour Avano. En effet certains moissonneurs embarquent des outils informatiques incapables de traiter les flux XML malformés. Ces moissonneurs sont donc incapables de traiter une archive comportant des problèmes d'encodage UTF-8 via la méthode *GetRecords*. En effet, un seul caractère corrompu dans un flux XML et ces moissonneurs ne sont plus capables de traiter les notices contenues dans le flux, n'y d'accéder aux notices suivantes en récupérant le *resumptionToken*.

Une solution de contournement devant ce type de problèmes consiste à moissonner les archives via la méthode *ListIdentifiers+GetRecord.* Dans ce cas, les notices qui contiennent des problèmes d'encodage ne sont pas intégrées dans la base de données du moissonneur, mais cette méthode permet au moins de moissonner la totalité de l'archive. Malheureusement ce n'est pas toujours possible, certaines archives, en effet, ne supportent pas la méthode *ListIdentifiers+GetRecord.*

Une autre solution consiste à tenter de contacter l'administrateur de l'archive incriminée pour lui signaler les erreurs détectées dans ses notices. C'est souvent la méthode la plus efficace car les administrateurs de ces archives sont souvent capables de corriger les problèmes rapidement. Mais tous ces contacts prennent du temps.

3.3. Moissonnage des grosses archives

Le moissonnage initial des très grosses archives et le moissonnage des archives particulièrement lentes sont également problématiques (les moissonnages suivants ne posent pas de problème si ces archives

supportent un moissonnage incrémentiel). En effet le moissonnage de ces archives peut nécessiter plusieurs jours, ou, pour quelques archives, plus d'une semaine. Si la moindre erreur survient lors du moissonnage de ces archives, il faut reprendre le processus depuis le début. En effet, le protocole OAI-PMH ne comporte pas de mécanisme de point de reprise qui permettrait au moissonneur de reprendre le traitement à partir de la dernière notice enregistrée.

Pour tenter de moissonner les grosses archives, ou celles qui sont à la fois lentes et instables, il est parfois possible de découper le traitement par plage de date, par exemple par année de mise à jour. Ce n'est malheureusement pas toujours possible :

- certaines archives ont mis à jour toutes leurs notices à un moment donné : dans ce cas, le traitement par plage de date ne permet pas de fractionner le traitement.
- Certaines archives offrent également de mauvaises surprises : si le traitement est découpé par tranche de dates, la somme de tous ces traitements ne permet pas de retrouver la totalité des notices contenues dans l'archive !

3.4. Gestion des doubles

Trop de doubles dans une liste de résultats nuit au confort des utilisateurs. Ce n'est pas aujourd'hui le principal problème auquel sont confronté les moissonneurs, mais ce problème risque de prendre plus d'importance dans les années à venir. Aujourd'hui, au moins deux phénomènes peuvent générer des doubles dans les bases de données des moissonneurs:

- Plusieurs organismes de recherche 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. Mais, vu le faible taux d'auto-archivage actuel, et particulièrement en sciences de la vie, ce phénomène n'est pas aujourd'hui le plus gros générateur de doubles.

- Les projets d'agrégateurs nationaux ou thématique sont plus problématiques. En effet, dans certains pays, des projets de fédération d'archives institutionnelles peuvent agréger les notices d'une sélection d'archives dans une base centralisée avant de les réexposer en OAI-PMH à partir de leur propre serveur. Les notices enregistrées dans ces serveurs sont donc exposées deux fois en OAI-PMH : via l'archive institutionnelle et via la base centralisée. Si l'administrateur d'un moissonneur ne maîtrise pas l'architecture de ces projets nationaux ou thématiques, il peut enregistrer ces différents serveurs et multiplier les doublons dans les listes de résultats de son moissonneur.

3.5. Gestion des notices supprimées

Certaines archives ne gardent pas la trace 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 peuvent proposer des fiches qui pointent vers des ressources qui n'existent plus. Pour contourner ce problème, les moissonneurs doivent re-moissonner complètement ces archives à intervalle régulier pour détecter d'éventuelles suppressions. Cette obligation est bien sur problématique pour les grosses archives ou les archives à la fois lentes et instables.

3.6. Gestion du champ Type

Pour respecter le protocole OAI-PMH les archives ont l'obligation d'exposer leurs données dans la DTD Dublin-Core non qualifiée. Dans cette DTD tous les champs sont optionnels. Ces champs sont également *non-qualifiés,* c'est-à-dire qu'ils ne sont pas tenus, par exemple, de correspondre à une liste fermée de valeurs. Ce caractère facultatif et non formalisé des informations pose plusieurs problèmes et notamment pour le champ *type*.

En effet, 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 textes libres (ex : « publication », « artjournal », « text », « article » sont utilisés pour décrire un article). Certains moissonneurs, dont Avano, proposent à leurs utilisateurs de limiter leurs recherches à un ou plusieurs types de ressources (Fig. 8). Pour permettre la mise en place de ce filtre les moissonneurs tentent de normaliser ce champ *type* à l'aide d'un système de reconnaissance de mots-clés dans cette chaîne de caractère. Cette normalisation n'est malheureusement pas parfaite, et ce

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. Certaines informations contenues dans ce champ titre sont d'ailleurs impossibles à normaliser. A titre d'exemple, la liste ci-dessous correspond à des champs types moissonnés par Avano :

- A1
- Airticle
- 8
- Treball Final de Carrera
- ...

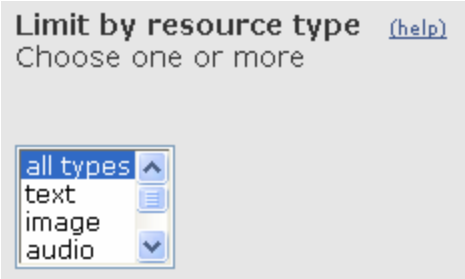

Fig. 8. Option proposée aux utilisateurs du moissonneur Oaister pour permettre de limiter une recherche à une sélection de type de données

Plus problématique encore, certaines archives ne renseignent pas ce champ. Dans le cadre d'Avano par exemple, plus de 26 000 notices sur les 107 000 disponibles en septembre 2007 ne présentent pas de champ *Type*. A moins de tenter de les renseigner manuellement (en contactant par exemple l'administrateur de l'archive pour vérifier avec lui que son archive ne contient des documents), toutes ces notices sont automatiquement exclues du périmètre de la recherche si un utilisateur limite sa requête à un ou plusieurs types sélectionnés.

3.7. Gestion du champ Date de publication

Le champ *Date de publication* pose les mêmes problèmes que le champ *Type*. Près de 15.000 notices (dont une majorité issues de PubMed Central) sur les 107.000 notices disponibles dans Avano en septembre 2007 ne disposent pas de date de publication. De plus, pour certaines notices, il n'est pas possible de normaliser ce champ *Date de publication*. C'est par exemple le cas pour les données suivantes moissonnées par Avano à partir de plusieurs archives:

- 1970-04-00
- 1981.
- Montréal, 2000
- [196-?]
- 2005-92-26
- ….

Quand une fiche ne contient pas de date de publication ou quand il n'est pas possible de la normaliser, elle se retrouve systématiquement en fin de liste quand l'utilisateur demande à trier sa liste de résultats par dates. De même, quand un utilisateur demande à limiter une recherche à une plage de dates spécifiques (Fig. 9), ces fiches sont exclues de la recherche même si elles correspondent à la recherche spécifiée.

Fig. 9. Option disponible dans le masque de recherche expert d'Avano pour permettre de limiter une recherche à une sélection de type de données

3.8. Notices pauvres

Certaines archives proposent des notices extrêmement dépouillées, avec par exemple un titre et un accès à l'objet numérique. Si l'objet numérique est un document, son enregistrement dans l'archive aura au moins le mérite de fournir un point d'accès aux robots des moteurs de recherche (ex. Google…). Mais pour les moissonneurs OAI, ces notices pauvres sont un vrai problème. En effet, la majorité (ou la totalité ?) des moissonneurs indexent uniquement la notice des documents. De plus la majorité des moissonneurs proposent, par défaut, une liste de résultats triée par *hit*, c'est-à-dire par nombre d'occurrences du mot recherché dans le texte. Dans les moissonneurs, ces notices pauvres auront donc une faible visibilité par rapport aux notices qui proposent un résumé du document.

3.9. Mélange de données brutes et de documentation

Aujourd'hui la grande majorité des archives disponibles propose principalement un accès à de la documentation. A l'avenir le contenu des archives pourrait se diversifier en proposant, par exemple, davantage d'images, de vidéos, de fichiers audio ou de jeux de données brutes.

De fait, plus de 90% des notices disponibles aujourd'hui dans Avano sont liées à de la documentation. Mais Avano propose également, par exemple, un accès à des banques d'image de plancton (http://planktonnet.eu/). Le mélange des notices liées à la documentation avec celles liées à ces images ne pose pas de problème et, au contraire, ce mélange documentation–image pourrait se révéler un des points forts des moissonneurs.

Par contre l'agrégation de notices liées à de la documentation avec des notices liées à des données brutes est plus problématique. En effet, les données, dans ces deux domaines, peuvent ne pas être proposées avec la même granularité. Le serveur Pangea est une bonne illustration de ce problème. Ce serveur propose un accès à des centaines de milliers de jeux de données brutes dans les domaines des géosciences et de l'environnement. Chacun de ces jeux de données est décrit par une notice accessible via le protocole OAI-PMH. Dans ce serveur, des milliers de notices peuvent ne différer que par leurs coordonnées géographiques. L'agrégation de cette archive avec des archives proposant de la documentation peut noyer une liste de résultats (Fig. 10).

La description en Dublin Core de ce type de données n'offre que peu d'intérêt pour des moissonneurs standard. Elle pourrait par contre être intéressante pour des moissonneurs spécialisés, capables d'offrir à leurs utilisateurs une interface de recherche graphique, si les notices étaient présentées dans une DTD capable de gérer des coordonnées graphiques de manière standardisée.

10. **Color scan of sediment core MD98-2162**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

11. **Color scan of sediment core MD98-2183**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

12. **Color scan of sediment core MD98-2190**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

13. **Color scan of sediment core MD98-2166**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

14. **Color scan of sediment core MD98-2163**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

15. **Color scan of sediment core MD98-2173**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

16. **Color scan of sediment core MD98-2191**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

17. **Color scan of sediment core MD98-2167**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

18. **Color scan of sediment core MD98-2192**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

19. **Color scan of sediment core MD98-2152**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-03-29, Dataset , Bassinot, Frank C

Fig. 10 : le site Pangea propose des milliers de notices bibliographiques quasiment identiques qui ne diffèrent, par exemple, que par une information relative aux coordonnées géographiques. Par exemple, ce serveur propose plus de 1000 notices quasiment identiques qui contiennent l'expression « color reflectance ». Si ces notices sont agrégées avec des notices liées à de la documentation, il sera impossible de retrouver les quelques fiches documentaires qui contiennent également cette expression dans ce millier de notices identiques.

3.10. Notices sans accès gratuit à l'objet numérique

Le protocole OAI-PMH définit uniquement un mécanisme de partage des notices bibliographiques contenu dans un ensemble d'archives. De fait, certaines archives mélangent aujourd'hui des notices sans lien vers l'objet numérique avec des notices permettant un accès gratuit à la ressource. D'autres archives proposent encore des notices avec un accès payant (ex : BePress) ou des notices avec un accès restreint, par exemple, au personnel d'une université.

C'est à mon sens le principal problème auquel sont confrontés les moissonneurs aujourd'hui. Il n'existe en effet aucune indication dans la DTD Dublin Core qui permette d'indiquer aux moissonneurs le degré d'accessibilité des objets que les notices décrivent. Les moissonneurs sont donc incapables de retranscrire cette information à leurs utilisateurs et encore moins d'offrir à leurs utilisateurs la possibilité de filtrer les notices vides ou les notices proposant un accès payant à la ressource.

A titre personnel, je pense d'ailleurs que la diffusion de notices sans texte intégral via une archive est aujourd'hui plus nocive qu'utile. Par manque de temps et/ou d'intérêt, les scientifiques peinent à adhérer au mouvement Open Access et le taux de dépôt en libre accès des publications reste marginal, notamment en science de la vie. Pour convaincre les scientifiques de l'intérêt du mouvement Open Access, l'accès gratuit et immédiat à la documentation est sans doute un des meilleurs arguments à faire valoir. Alors noyer une minorité de notices proposant des publications en libre accès sous des ensembles de notices sans lien vers le texte intégral et/ou des notices offrant un accès payant aux documents n'est peut-être pas la meilleure solution aujourd'hui pour promouvoir le mouvement Open Access.
Encore une fois, ces notices sans accès gratuit au texte intégral ne seraient pas un problème pour les moissonneurs si la DTD Dublin Core permettait d'indiquer aux moissonneurs le degré d'accessibilité des objets que les notices décrivent. Les moissonneurs pourraient alors offrir à leurs utilisateurs la possibilité de filtrer les notices sans accès gratuit à l'objet numérique. Mais ce n'est pas le cas aujourd'hui.

3.11. Moissonnage thématique

Si le moissonnage thématique d'une archive est supposé possible via le mécanisme des *Set* du protocole OAI-PMH, dans la réalité nous n'avons bien sûr jamais trouvé de *Set* «sciences marines et aquatiques» dans aucune des archives que nous avons moissonnées. Ce mécanisme de *Set* est optionnel dans le protocole OAI-PMH et, dans la réalité, en l'absence de recommandation, il est implémenté de manière très diversifiée. Certaines archives proposent effectivement des *Set* thématiques qui correspondent parfois à un plan de classement hérité du classement de leur collection papier. Certaines archives proposent également des *Set* par type de document (publications, rapports, thèses…) ou en fonction de l'état du document (publications InPress, publiées,...). D'autres proposent encore des *Set* qui permettent d'isoler les notices qui permettent un accès à l'objet numérique si l'archive contient également des notices vides.

La mise en place d'un plan de classement thématique pourrait sans doute s'envisager au sein d'une petite communauté d'organismes scientifiques. Mais au plan international, il est certainement impossible d'accorder la communauté scientifique mondiale sur un plan de classement thématique unique. C'est pourquoi nous avons développé ce mécanisme de filtre des notices basé sur la recherche de mots-clés. Cela nous semble la seule méthode réaliste pour mettre en place un moissonneur thématique sur l'ensemble des archives disponibles au niveau mondial.

4. Conclusion

Nous l'avons vu, même s'il augmente, le nombre de connexions à Avano reste aujourd'hui relativement faible. Nous pouvons sans doute l'expliquer par le fait que:

- Tous les scientifiques et tous les étudiants ont accès à Google/Google Scholar avec ses milliards de pages indexées. Les scientifiques ont également accès à un ensemble d'Archives Ouvertes devenues incontournables dans leurs domaines (ex : ArXiv, PubMed Central…). Nous pouvons espérer également qu'une grande majorité des scientifiques, au moins dans les pays occidentaux, ont accès à des bases de données commerciales de références (e.g. Web Of Science, Scopus, …) qui couvrent la majorité de la production scientifique mondiale. A coté de Google, qui référence lui-même le texte intégral de la très grande majorité des documents

référencés par Avano, et à coté des bases de données bibliographiques commerciales, les moissonneurs ne référencent qu'une partie marginale de la production scientifique mondiale du fait du faible taux de dépôt en libre accès des publications, et notamment en science de la vie.

- L'ensemble des moissonneurs, et ils sont de plus en plus nombreux (ex : Oaister, BASE, CyberThèses, Avano, Socolar, Scientific Commons …), se partage un même public en offrant tous l'accès aux mêmes documents.

- Un moissonneur n'a pas de contenu propre qui lui permette de gagner en visibilité sur le WEB. Pour se faire connaître, il ne peut compter que sur des opérations de promotion classiques (email sur des listes de diffusion, référencement sur des portails thématiques). A titre de comparaison, le site WEB Archimer, l'archive institutionnelle de l'Ifremer, qui ne contient que 2300 documents en texte intégral est plus visité que le site d'Avano avec ces 100.000 références. En effet, chaque nouveau document enregistré dans Archimer est indexé par Google. Il devient donc une nouvelle porte d'accès à Archimer ouverte sur le WEB. En effet, si un lecteur découvre un document enregistré dans Archimer via Google (90% des documents enregistrés dans Archimer sont déchargés via Google), et si le document l'intéresse, il rebondit sur le site d'Archimer par le lien posé à cet effet sur le texte intégral.

- Par rapport aux moteurs de recherches standard du Web (ex : Google), les moissonneurs auraient dû pouvoir offrir des options de recherche plus perfectionnées, comme par exemple, des recherches sur la date de publication. Or, comme nous l'avons mentionné, la mauvaise qualité des données bibliographiques proposées par certaines archives dégrade les services proposés par les moissonneurs.

- Par rapport aux bases de données bibliographiques commerciales, les moissonneurs auraient dû pouvoir se prévaloir d'offrir un accès systématique et gratuit à l'ensemble des objets numériques et notamment à la documentation. Mais nous l'avons vu, de plus en plus d'archives noient une minorité de notices proposant des publications en libre accès sous des ensembles de notices sans lien vers le texte intégral et/ou des notices offrant un accès payant aux documents.

Alors, que faudrait-il aux moissonneurs pour qu'ils puissent trouver leur public? Les améliorations suivantes pourraient sans doute les aider :

- Une augmentation sensible du taux de dépôts des publications en sciences de vie permettrait d'obtenir une masse critique de documents accessibles gratuitement.

- Une adoption du protocole OAI-PMH par davantage d'éditeurs commerciaux permettrait également de couvrir rapidement la plus grande proportion possible de la production scientifique mondiale.

- Une amélioration de la version actuelle du protocole OAI-PMH pour permettre un moissonnage plus facile des archives et pour garantir davantage de qualité dans les notices. Cette amélioration du protocole OAI-PMH pourrait, par exemple, porter sur :

    - Un système de reprise qui permette aux moissonneurs de reprendre un traitement interrompu par une erreur à partir de la dernière notice enregistrée.
    - L'ajout de champs obligatoires et normalisés, et notamment portant sur les champs *Date* et *Type*,
    - L'ajout d'une information obligatoire et normalisée, au niveau de la notice, indiquant le degré d'accessibilité de l'objet numérique (gratuit, payant, impossible, restreint…),
    - L'ajout d'une information dans la description de l'archive portant sur la participation ou non de l'archive à un système d'agrégation national ou thématique qui réexposerait les notices en OAI-PMH à partir d'un autre serveur. Cette information devrait permettre aux administrateurs des moissonneurs OAI d'éviter d'enregistrer des archives qui n'apportent que des doublons.
    - …

- …

Et tous les cas, l'actuelle faiblesse des connexions à Avano ne remet pas en cause l'intérêt des archives ouvertes. Les autres moissonneurs, et notamment ceux qui disposent de gros moyens de développement, sont certainement plus utilisés qu'Avano. Et surtout, les documents enregistrés dans les archives sont souvent consultés massivement. Ils ne sont tout simplement pas consultés via le site WEB des archives institutionnelles, ni via les moissonneurs, mais par Google.

Enfin, si la première Archive Ouverte date de plus de 15 ans, si le protocole OAI-PMH date de 6 ans, le mouvement Open Access ne semble vraiment émerger que depuis l'année 2006, au moins dans le domaine des sciences de la vie. Nous pouvons donc espérer que le taux de dépôt des publications en libre accès continuera à augmenter pour arriver à une part significative de la production scientifique mondiale. Nous pouvons également espérer qu'une future version du protocole OAI-PMH donnera aux moissonneurs les moyens d'offrir un service de qualité à ses utilisateurs.

Le principal intérêt des moissonneurs pourrait se limiter aujourd'hui à participer à la promotion du mouvement Open Access, et ce n'est pas rien. Demain ils pourraient offrir une réelle utilité et trouver leur public.

5. Références

Timothy W. Cole. What Is OAI-PMH Good For?
http://www.sis.pitt.edu/~egyptdlw/papers/Timothy_Cole.html

Muriel Foulonneau, Friedrich Summann, Jochen Schirrwagen, Paul Walk, Dr. Peter Millington, Laurents Sesink, Maarten Steenhuis, Kasper Løvschall, Franck Falcoz (Feb. 2007). Institutional Repositories Workshop Strand Report Strand title: Open Archives Protocol for Metadata Harvesting
http://www.knowledge-exchange.info/Default.aspx?ID=164

The Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/OAI/openarchivesprotocol.html