L'archive ouverte HAL

Daniel Charnay, Christian Michau

L'archive ouverte HAL

Daniel CHARNAY
CNRS - CCSD
27, Boulevard du 11 novembre, 69622 Villeurbanne
charnay@in2p3.fr

Christian MICHAU
AMUE
103, Boulevard Saint Michel, 75005 Paris
Christian.Michau@amue.fr

Résumé

Cet article présente, avec une approche essentiellement technique, la plate-forme commune inter-établissements d'archives ouvertes, HAL. Après une courte introduction au concept d'archives ouvertes et aux technologies de ce type de serveur, les principales fonctionnalités du serveur HAL seront décrites. Cette description devrait permettre de mieux comprendre comment un établissement peut s'approprier l'archive en faisant le choix d'un hébergement complet ou l'insérer le plus efficacement possible dans son propre système d'information. L'examen des protocoles supportés, qu'ils soient dédiés comme OAI-PMH ou plus généraux comme SOAP, sera abordé ainsi que l'aspect intégration dans un système d'information local.

Mots clefs

HAL, Archives ouvertes, Open Access, OAI, OAI-PMH, ArXiv, PubMed, Sherpa/Romeo, REDIF, REPEC, Dublin Core, COST, AO.fr

1 Introduction

Dans le cadre du mouvement mondial en faveur du libre accès se constituent dans le monde des réservoirs thématiques de la production scientifique « académique » sur le mode des archives ouvertes. Les institutions de recherche cherchent à la fois à pérenniser leur production scientifique tout en leur donnant un maximum de visibilité au sein des communautés internationales. Ce modèle de communication scientifique directe, qui se situe en parallèle au circuit traditionnel des revues scientifiques, est fondé sur l'auto-archivage par les chercheurs de leurs articles dans un contexte où un nombre croissant d'éditeurs acceptent le dépôt des versions auteur sur ces archives ouvertes.

L'archive ouverte HAL[1] créée en 2000 à l'initiative du CNRS et développée par le CCSD[2], est destinée au dépôt et à la diffusion des travaux de recherche (articles, thèses, actes de conférences, etc.), de toutes les disciplines scientifiques.

Elle est couplée à des archives internationales comme ArXiv[3] (physique, mathématiques, informatique) et Pubmed Central[4] (sciences de la vie). Cette bibliothèque numérique contribue à la libre diffusion (internationale, immédiate, gratuite) et à la valorisation du savoir scientifique produit par les chercheurs. HAL constitue donc un élément majeur pour favoriser l'accès au savoir et l'essor de la science, pour les chercheurs du monde entier et particulièrement des pays en voie de développement, face aux budgets importants qu'impliquent les abonnements aux revues scientifiques [1].

La signature, en juillet 2006, à l'académie des sciences, d'un protocole d'accord entre le CNRS, l'INRIA, l'INRA , l'INSERM, le CEMAGREF, l'IRD, l'institut Pasteur, la CPU, la CGE et plus récemment par le CEA, l'IFREMER, l'INERIS, l'INRETS positionne Hal au cœur de l'archive ouverte de la recherche française.[5]

2 Archive ouverte : principes généraux

Fonctionnellement il s'agit de fournir un site Web sur lequel un contributeur précédemment identifié (souvent un des auteurs de l'article) viendra télécharger (*upload*) le texte intégral du document (publication, thèse, etc.) assorti d'un certain nombre de métadonnées identifiant, classifiant son dépôt. Ces métadonnées contribueront, entre autres, à faciliter la recherche pour l'internaute qui viendra, de façon anonyme, sur ce même site pour consulter des textes (*download*). Une administration du site permettra le contrôle des dépôts, celui-ci consistant essentiellement en une validation scientifique élémentaire assortie d'un contrôle de la qualité du fichier déposé (format, lisibilité, accessibilité, etc.). Outre ces accès via une interface

[1] Hyper Articles en Ligne, http://hal.archives-ouvertes.fr

[2] Centre pour la Communication Scientifique Directe. http://ccsd.cnrs.fr

[3] La création par Paul Ginsparg en 1991 à Los Alamos de l'archive http://arxiv.org est l'origine même des archives ouvertes. Le miroir français d'ArXiv, http://fr.arxiv.org/ est hébergé au CCSD.

[4] http://www.pubmedcentral.nih.gov/

[5] Le protocole d'accord, les textes fondateurs sont disponibles sur http://www.archives-ouvertes.fr.

« humaine », une archive ouverte devra exposer son contenu à des systèmes automatiques qui pourront exploiter les métadonnées aux fins de constitutions de grands catalogues mondiaux (Oaister, Google scholar, etc.). Cette mise à disposition des métadonnées de chacun des articles est réalisée grâce à un protocole extrêmement simple constitué de quelques verbes encapsulés dans des requêtes HTTP ; il s'agit du protocole OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*)[6].

3 Technologies de base

Les composants essentiels d'une archive ouverte sont donc un serveur HTTP qui va héberger les différents formulaires de dépôt, de consultation, d'administration, le module qui répondra aux requêtes OAI-PMH, une base de données qui va recueillir les métadonnées propres à chacun des articles et un espace de stockage qui va contenir les fichiers du texte intégral des articles. On peut ajouter à ces quelques composants un moteur d'indexation du texte intégral qui permettra un autre moyen de recherche que celui fourni au travers des métadonnées. Ces composants, généralement issus du logiciel libre, sont souvent à base de PHP, PERL ou JAVA pour la fabrication des formulaires qui s'exécutent sur une plate-forme de type Apache hébergée le plus fréquemment sur des serveurs Linux, tandis que pour la base de données on trouvera souvent MySql ou PostgreSQL. Le moteur d'indexation associé pouvant être du type htDig[7].

Il existe maintenant un certain nombre de logiciels d'archives ouvertes[8] pouvant être déployés facilement comme ePrint ou dSpace. Un groupe d'établissements universitaires a développé une approche fédérative qui est mise en œuvre dans le logiciel ORI/OAI[9].

4 La problématique des métadonnées

Ces informations qui vont permettre de décrire la ressource elle-même (l'article en texte intégral) sont renseignées par le contributeur lors du dépôt de son article ; elles devront donc être les plus concises possibles afin de ne pas décourager le déposant par un nombre de champs à remplir trop important. Ce minimum de métadonnées est souvent une déclinaison du Dublin Core[10] non qualifié. En revanche, de la qualité de ces métadonnées va dépendre l'exploitation qui pourra être faite de l'archive ouverte. Ainsi, par exemple, si on prend soin de recueillir pour chaque auteur d'un article l'ensemble de ses affiliations (laboratoire et ses tutelles), il sera extrêmement facile d'extraire de l'archive ouverte toutes les vues institutionnelles souhaitées[11]. Dans cet exemple on voit que la constitution des archives institutionnelles ne fonctionnera bien que si l'affiliation est issue d'un référentiel fixe et non d'une saisie dans un champ libre ; on ne doit en effet trouver qu'une seule occurrence d'une entité donnée. (CEA et non pas C-E-A, C E A, ou C.E.A)

Dans une archive ouverte, chaque fois que cela est possible, les métadonnées seront choisies dans des référentiels fixes et non saisies dans des champs libres. Si des archives ouvertes doivent inter-opérer entre elles, elles devraient obligatoirement partager les mêmes référentiels, ceux-ci pouvant être indépendants du contexte d'archives ouvertes (Le référentiel des unités de recherche et de leurs affiliations devrait être fourni à terme par le ministère de la recherche et sera utilisé par bien d'autres applications).

5 L'archive ouverte HAL

Si le système informatique est centralisé on peut considérer que HAL est une archive mutualisée permettant de créer des entrées spécifiques ou portails, à caractère institutionnel ou disciplinaire. (Ces portails sont parfois abusivement appelés « instances »). Sa technologie est du type LAMP (Linux-Apache-MySql-PHP). Lors de sa création en 2000 il n'existait que la plate-forme ePrint comme logiciel d'archive ouverte, mais elle ne répondait pas au cahier des charges du CCSD ; il a donc été décidé de développer entièrement un logiciel d'archives ouvertes.

L'unité documentaire dans HAL

L'unité documentaire dans HAL est constituée d'une ou plusieurs versions d'un même article comportant chacune une notice descriptive (les métadonnées) incluant, si le document est publié, ses références bibliographiques, un ou plusieurs formats du même fichier dont un format accessible depuis un logiciel gratuit (PDF, PostScript, HTML, etc.). Il est possible d'adjoindre au document primaire des documents annexes (par exemple le diaporama de la soutenance d'une thèse, une animation 3D représentant une molécule, une vidéo, etc.) Cette unité documentaire est « citable » électroniquement à partir d'une URL simplifiée dont la pérennité est assurée (Exemple http://hal.archives-ouvertes.fr/hal-00157941/fr/). L'ensemble des versions d'un même article est accessible au lecteur.

Les fonctionnalités de HAL

Nous allons décrire brièvement les principales fonctionnalités de HAL essentiellement du point de vue de l'utilisateur de l'archive. Nous n'examinerons pas dans cet article tous les rôles particuliers permettant la gestion des dépôts, des interfaces, etc. Il faut simplement savoir qu'un certain nombre de rôles attribuables aux laboratoires ou

---

6 http://www.openarchives.org/OAI/openarchivesprotocol.html

7 http://www.htdig.org

8 http://www.eprints.org/, http://dspace.mit.edu/, http://www.fedora-commons.org/

9 http://www.ori-oai.org/

10 Il s'agit d'une normalisation d'un ensemble simple de 15 définitions permettant de décrire des ressources sur Internet. (norme ISO 15836 depuis février 2003).

11 Par exemple si l'auteur Jean Dupont est affilié au laboratoire Delta et que ce laboratoire est référencé comme une unité mixte entre le CNRS, l'INSERM et l'université de Lyon 1, alors sa publication enrichira automatiquement les vues institutionnelles de Delta, CNRS, INSERM et université de Lyon 1.

aux établissements vont permettre de corriger ou de compléter les métadonnées d'un dépôt, ou d'intervenir sur le style d'un portail ou d'une collection.

Les portails

HAL permet la création de portails (*Figure 1*), qu'ils soient institutionnels, disciplinaires, ou liés à un type de document particulier comme les thèses reçues sur un portail spécifique de HAL, TEL. Pour faire un raccourci on pourrait dire que dans la majorité des cas, un portail est un sous-ensemble du portail générique hal.archives-ouvertes.fr. Un portail est donc une entrée particulière permettant à la fois le dépôt et la consultation. Tous les articles déposés dans un portail sont forcément accessibles depuis l'entrée générique HAL. Le portail peut avoir une adresse internet dans le domaine de son choix et être personnalisé aux couleurs de l'institution qui a demandé sa mise en place ; le gestionnaire du portail peut agir sur le graphisme et la présentation de la page d'accueil au travers d'outils mis à sa disposition et à l'aide de feuilles de style CSS. Au-delà du style du portail, celui-ci peut ne présenter que le sous-ensemble des disciplines qui le concerne. Par exemple hal.inria.fr proposera essentiellement l'informatique, les mathématiques, et quelques autres domaines pertinents dans ses domaines de recherche. Un portail peut aussi étendre le modèle de données en ajoutant quelques métadonnées qui lui sont spécifiques (un thésaurus, une référence particulière, etc.). Un portail peut décider de montrer uniquement les dépôts qui sont faits au travers de ce portail, ou au contraire choisir de substituer à cette vue une vue « collection ». Par exemple le portail sciences de l'environnement va montrer tous les articles pertinents dans ses domaines, ceux-ci ayant pu être déposés depuis le portail générique ou tout autre portail.

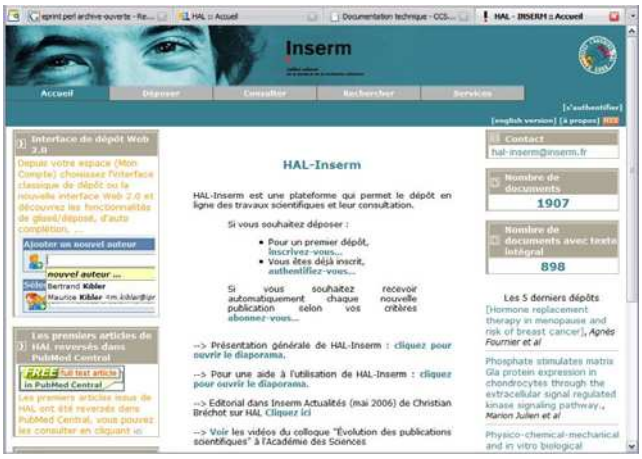

Figure 1 - *Le portail INSERM*

Les collections

HAL implémente le concept de collection. Une collection est le regroupement d'un ensemble d'articles sous une URL particulière du type hal.archives-ouvertes.fr/ COLLECTION/langue et pouvant être personnalisée par le propriétaire de la collection. Le graphisme, les textes de la page d'accueil de la collection peuvent être gérés directement depuis une interface Web mise à la disposition du propriétaire de collection. Une collection se constitue en apposant sur des articles le tampon de la collection. Cette métadonnée peut être ajoutée manuellement ou automatiquement sur un critère spécifique. Par exemple la collection regroupant tous les articles dont au moins un auteur appartient à un laboratoire affilié à l'INSU va créer la collection INSU, l'archive institutionnelle de cet institut. Les collections peuvent donc servir à constituer les actes d'un congrès, un journal en ligne, le présentoir des publications du laboratoire, etc. Chaque article peut appartenir à autant de collections que nécessaire, l'article ayant pu être déposé depuis n'importe quel portail de HAL. Du point de vue OAI-PMH une collection va générer un *set* spécifique. Une extension des collections permet de générer des ensembles de collections, par exemple, la collection d'un journal sera constituée par des collections représentant chacune un numéro du journal. Ce numéro est aussi une collection qui regroupe tous les articles du numéro.

Les référentiels

La qualification des métadonnées est fondée sur un certain nombre de référentiels qui sont actuellement tous hébergés au CCSd. En dehors de l'application HAL ils sont accessibles via des web services (SOAP). Les référentiels les plus importants sont les suivants :

- Le référentiel des laboratoires : c'est au niveau laboratoire que le déposant va affilier les auteurs de la publication. L'enregistrement d'un laboratoire contient le nom du laboratoire, son sigle, l'ensemble des institutions, des universités, et des grandes écoles auquel il est rattaché, l'URL du laboratoire, son adresse. Il existe aussi une information indiquant si le laboratoire est en activité ou s'il est fermé. HAL utilisera le référentiel national sitôt qu'il sera disponible.

- Le référentiel des domaines scientifiques : une publication est obligatoirement attachée à un domaine scientifique principal et peut être rattachée à autant de domaine secondaire que nécessaire. Ceci permet de caractériser la transdisciplinarité de la publication.

- Le référentiel des revues : il permet le choix entre 25 000 titres ; les éléments sont les informations classiques des revues, titre, éditeur, ISSN, racine du DOI[12], etc. Ces éléments sont complétés par les informations relatives à la position de l'éditeur

[12] Système permettant de donner un identifiant numérique à un document indépendant de son emplacement physique. Utilisé par les éditeurs scientifiques pour donner accès aux copies numériques des articles des revues. http://www.doi.org/

vis-à-vis des archives ouvertes ; elles sont issues de Sherpa/Romeo[13] et mises à jour de façon hebdomadaire. Cette information est restituée au déposant, à titre indicatif, lors de la sélection d'une revue.

L'interface de dépôt

L'utilisateur peut choisir dans ses préférences deux modes différents. Des formulaires classiques ou une interface plus élaborée, à base d'Ajax (*Figure 2*), avec des champs en auto-complétion, des possibilités de glissé-déposé comme pour reclasser l'ordre des auteurs par exemple, etc.

Figure 2 - *Interface Ajax*

Il est fréquent que les informations saisies par le contributeur préexistent dans une autre base, surtout si l'article a été précédemment publié. HAL reconnaît un certain nombre de base de données comme Pubmed, ADS (NASA), etc. ou des bases institutionnelles comme celle de l'IRD par exemple. En saisissant l'identifiant dans une de ces bases HAL collectera le maximum de métadonnées et pré-remplira tous les champs possibles. Le COST, groupe de travail de l'accord inter-établissements a défini un document où est décrit un format OAI spécifique AO_FR que chacun des signataires s'engage à implémenter s'il décide d'avoir une archive locale.

Le dépôt par transfert de fichiers

Utilisé essentiellement pour transférer des lots de documents en une seule opération, par exemple pour migrer une base locale, HAL permet le transfert d'un fichier contenant un « train » de documents. Ce fichier est construit en XML selon un schéma défini ; il contient une succession d'ensembles métadonnées avec les fichiers encodés en base64 ; il peut ainsi être téléchargé dans l'archive. A réception le fichier XML est analysé et s'il est correctement formé et si les données sont validées par le schéma XML, les documents sont candidats pour être validés puis insérés le cas échéant dans l'archive.

L'indexation en texte intégral

L'ensemble des documents contenu dans l'archive HAL est indexé afin de permettre la recherche en texte intégral, Le CCSD a acquis une « *appliance* » Google (boite noire intégrant le matériel et le logiciel du moteur) ; d'autres types de moteurs comme un moteur sémantique seront sans doute mis en œuvre dans le futur.

Les protocoles supportés dans HAL

OAI-PMH

Une archive ouverte implémente obligatoirement le protocole d'interopérabilité OAI-PMH[2] ; HAL est conforme à ce protocole dans sa version 2.0. Pour mémoire ce protocole a pour objectif la mise à disposition libre des descriptions des documents contenus dans une archive ouverte (les notices). Le transfert des données est assuré via le protocole HTTP et les données sont encapsulées en XML selon des schémas XML publiés et accessibles. L'OAI ne définit pas le contenu des métadonnées des notices mais spécifie un format minimum afin d'assurer un dénominateur commun entre des archives ouvertes hétérogènes. Ce minimum retenu est le Dublin Core non qualifié. Au sens OAI des acteurs du protocole, HAL est un entrepôt de données et fournit donc le metadataFormat OAI_DC, mais il en fournit aussi d'autres dont le format OAI_HAL plus complet et plus proche du Dublin Core qualifié, et plus récemment le format OAI_AOFR défini par le comité inter-établissements. Ce format sera la base commune utilisée pour toutes les interactions entre les archives du domaine archives-ouvertes.fr. Les principaux fournisseurs de service qui moissonnent actuellement la base HAL (OAISTER, GOOGLE SCHOLAR) ne vont pas au delà du format OAI_DC. Il est sans doute important de rappeler qu'OAI-PMH est un protocole visant la constitution de catalogue et que la fréquence du moissonnage, donc l'actualisation des données, dépend du fournisseur de service. Ce n'est donc pas un protocole destiné à la recherche multi-bases en temps réel comme peuvent l'être Z39.50 ou SRU/W [3]

ReDIF[14] (Research Documents Information Format).

Le concept, antérieur à OAI-PMH, a été défini pour une base de référence des articles en économie, RePEc[15]. Cette base implémente des agents qui collectent sur des serveurs qui lui sont connus les notices d'articles aussi encapsulées en XML. Peu de logiciels d'archives ouvertes implémentent ce format pourtant indispensable aux chercheurs en économie.

Les Web Services - SOAP[16]

Non l'avons dit, HAL est un fournisseur de données et n'a aucune vocation à devenir un moissonneur. Dans un contexte d'interconnexion d'applications hétérogènes, les archives ouvertes ou les bases documentaires étant distribuées sur différentes plates-formes allant de celles qui

---

13 http://www.sherpa.ac.uk/romeo.php

14 http://openlib.org/acmes/root/docu/redif_1.html

15 http://www.repec.org

16 http://www.w3.org/TR/soap/

utilisent un logiciel libre, à d'autres implémentées sur des logiciels commerciaux, HAL a été sollicité pour que l'on puisse déposer des documents dans l'archive sans utiliser l'IHM. En d'autres termes certains établissements disposent d'une base de données documentaire et souhaitent qu'un dépôt dans leur base alimente automatiquement la plate-forme commune, HAL. La mise à disposition de Web Services est la réponse apportée par HAL[17]. Les Web Services de HAL utilisent le protocole SOAP (Simple Object Access Protocol) et le protocole HTTP pour le transport. Ce n'est pas comme les deux protocoles cités précédemment, un protocole d'interopérabilité dédié aux archives ouvertes, mais un protocole généraliste type RPC, qui au travers d'échange de messages exprimés en XML permet d'invoquer des fonctions ou méthodes sur un serveur distant. HAL fournit donc des méthodes pour accéder aux référentiels, des méthodes pour déposer ou modifier un dépôt, ainsi qu'une méthode pour rechercher dans l'archive. La description de l'interface d'utilisation de ces méthodes est directement accessible depuis le WSDL généré (Web Service description Language).

Le scénario de dépôt d'une archive locale vers HAL peut donc se schématiser ainsi : le client acquiert les référentiels de HAL ; il peut le faire au coup par coup mais généralement il les trouvera dans un cache qu'il aura constitué et régulièrement actualisé. Il qualifiera les données en accord avec ces référentiels pour ensuite effectuer le dépôt en invoquant la méthode adéquate. HAL lui répondra de façon synchrone en lui retournant un identifiant. Le document entrera alors dans le processus de validation scientifique sommaire, cette validation étant effectuée par des personnes, c'est de façon asynchrone que le client recevra l'information de mise en ligne. Ce client devra donc fournir aussi un Web Service qui recevra l'identifiant de l'article et le statut de mise en ligne (dépôt accepté ou refusé).

Les modèles d'interconnexion entre un système local d'établissement et HAL

On désignera par *Système Local* tous les systèmes pouvant être mis en place dans les établissements : systèmes d'information, systèmes documentaires, archives institutionnelles…

Les Systèmes Locaux ont en général des finalités supplémentaires de suivi des publications, voire d'évaluation ou de production d'indicateurs. Ils peuvent contenir aussi bien des notices que des documents numériques. L'ensemble des métadonnées qualifiant ces documents peut être à la fois plus important et plus spécifique à l'établissement, mais devra intégrer obligatoirement le cœur commun des métadonnées défini dans HAL ; l'utilisation des référentiels communs est donc une obligation.

Les Systèmes Locaux relèvent de la politique des établissements et répondent à des objectifs de valorisation et de diffusion des publications propres à un établissement ou/et à un réseau d'établissements, ne relevant pas forcément du périmètre de HAL (rapports, cours, mémoires d'étudiants, etc.). A ce titre les Systèmes Locaux doivent être interopérables avec d'autres archives institutionnelles, thématiques, nationales et internationales.

L'interopérabilité s'organise à travers trois dispositifs complémentaires :

- des règles d'interopérabilité régissant les situations où un document a vocation à être déposé dans un système local et dans HAL.
- un ensemble de métadonnées – appelé AO.fr – permettant des échanges de données bibliographiques et aussi de gérer la cohérence des informations entre les différents dépôts. Ce format de métadonnées sera aussi utilisé pour le pré remplissage des champs de formulaires quand l'information est présente sur le système local.
- les Web services disponibles dans HAL pour importer ou exporter des documents avec un système local ou permettre à un système local d'extraire des informations de HAL en temps réel.

Le document de spécification pour l'interopérabilité est en cours de finalisation par un sous-groupe du COST[18]. Dès qu'il sera disponible il sera publié sur le site du CCSD dans la rubrique documentation technique.

Le couplage avec les archives internationales

Du point de vue du chercheur et pour certaines disciplines scientifiques, la publication d'un article hors de certaines bases internationales est inenvisageable tout autant que le double dépôt. Il était donc indispensable que le serveur HAL propage les articles vers ces sites mondiaux avec la même rapidité de mise en ligne qu'un dépôt direct. Ces contraintes ont amené HAL à établir un couplage particulier avec des règles négociées directement avec les opérateurs de ces archives.

Le couplage avec ArXiv

HAL est donc un point d'entrée privilégié vers la base d'ArXiv (domaines de physique, mathématiques, informatique essentiellement). D'un point de vue informatique le protocole entre HAL et ArXiv a été établi entre les deux opérateurs et n'utilise pas de standards d'interopérabilité.

---

[17] Consulter la documentation technique sur http://ccsd.cnrs.fr rubrique documentation, documentation technique.

[18] Comité Scientifique et Technique du protocole « archives-ouvertes.fr », http://www.archives-ouvertes.fr rubrique COST

L'interconnexion avec PubMed Central

Cette archive est la référence dans les sciences de la vie. La problématique du transfert vers cette archive est techniquement encore plus lourde. La base de PMC stocke les articles en XML, il est donc nécessaire d'effectuer une transformation du fichier vers ce format selon une DTD très stricte. PMC préconise une transformation manuelle effectuée depuis la version PDF de l'article. Le *workflow* d'un article déposé dans HAL et à destination de PMC est le suivant :

- Dépôt dans HAL avec importation des métadonnées depuis Pubmed[19] (l'article à destination de PMC est forcément publié)
- Réception dans HAL, mise en ligne le cas échéant
- Transfert de la version PDF vers une société de service qui va procéder à la conversion en XML
- Retour dans HAL du document XML et d'une version PDF générée depuis le XML produit
- Transfert vers le déposant du nouveau PDF pour validation
- Si la vue PDF du document XML est validée par le déposant, transfert vers PMC sinon demande de corrections à la société de service.

Si parfois il peut paraître trivial et valorisant de créer sa propre archive ouverte locale à partir d'un logiciel libre, l'interconnecter avec les bases mondiales de référence peut s'avérer difficile, coûteux, voir impossible lorsque l'opérateur fait le choix d'une seule connexion par pays. Pour les archives françaises qui lui sont reliées, HAL pourra remplir le rôle de relais vers l'international dans la mesure où les contraintes techniques le permettront.

L'interconnexion avec les applications administratives

L'objectif du dépôt unique doit impérativement être recherché. Convaincre le chercheur de déposer dans une archive ouverte et expliquer une démarche qui doit rester scientifique et l'assurer que les informations administratives seront extraites automatiquement contribuera au succès de l'archive.

Actuellement quelques applications institutionnelles s'alimentent directement à partir de HAL ; citons, pour le CNRS, Labintel production (référentiel des publications du CNRS) et CRAC (compte-rendu Annuel d'Activité Chercheurs dont la liste des publications est pré-remplie si le chercheur a déposé des publications dans HAL). Une démarche identique via l'application GRAAL est actuellement mise en place par un consortium universitaire.

Les services annexes pour le chercheur, le laboratoire, etc.

Outre l'accès au texte intégral d'environ 50 000 articles[20] scientifiques, HAL fournit un certain nombre de services comme les fils RSS, les abonnements, et permet la génération de listes de publications formatées à la demande et dans des formats divers (TeX, RTF, XML, PDF, etc.). Pour les publications écrites en LaTeX, HAL fournit la compilation en ligne et l'accès au compilateur pour valider le document. Les contributeurs de l'archive ont accès aux statistiques de téléchargement des articles qu'ils ont déposés.

Enfin citons un certain nombre d'applications commerciales comme des services de gestion de congrès, des logiciels de gestion de bibliothèque ou des gestionnaires de bibliographie qui proposent une interconnexion avec HAL.

La participation aux projets européens

Depuis juin 2006, le CCSd participe au projet européen DRIVER [4] (Digital Repositories Infrastructure Vision for European Research)[21]. DRIVER crée une infrastructure de recherche pour créer un espace numérique pour les contenus scientifiques européens. Une architecture orientée services (SOA) est construite pour mutualiser des services d'indexation, d'authentification, de retraitement de données. Des recommandations communes sont élaborées sur les technologies et standards pour les archives (implémentation OAI-PMH, métadonnées, représentation d'objets complexes). Le projet DRIVER assure l'interconnexion des archives européennes et doit lancer la création d'une fédération européenne des archives scientifiques.

Conclusion

La plate-forme commune HAL est une solution immédiate pour les établissements qui souhaitent participer au mouvement des archives ouvertes et permet, à ceux qui font le choix de construire un système d'information plus global, de s'interconnecter grâce à des solutions définies par les instances communes.

HAL assure une visibilité internationale aux contenus scientifiques en s'interconnectant elle-même avec des

[19] PubmedCentral et Pubmed sont des bases différentes. Pubmed ne recense que les notices bibliographiques des articles publiés alors que PMC qui est une archive ouverte en diffuse le texte intégral.

[20] Novembre 2007. Le nombre de dépôts mensuel se situe à cette date autour de 1500 articles par mois, ce qui représente très approximativement entre 15 et 20% de la production scientifique française.

[21] http://www.driver-support.eu

archives internationales telles que PubMed Central et arXiv. HAL participe à la construction des infrastructures numériques de recherche au travers du projet européen DRIVER (Digital Repository Infrastructure Vision for European Research). HAL assure ainsi une intégration technologique dans les réseaux scientifiques internationaux.

Bibliographie

1-Francis André, "*Libre accès au savoir*", page 19, Futuribles : Paris 2005

2-Muriel Foulonneau, "Développements du protocole OAI-PMH dans le monde", pages 63-94, et "Assurer l'interopérabilité des systèmes documentaires", pages 163-185 "*Les Archives Ouvertes, enjeux et pratiques*", dir. Christine Aubry et Joanna Janik, ADBS : Paris, 2005.

3-Timothy W. Cole, Muriel Foulonneau, "*Using the Open Archives Initiative Protocol for Metadata Harvesting*" Westport, CT : Libraries Unlimited, 2007.

4-Francis André, Muriel Foulonneau, Anne-Marie Badolato, and Daniel Charnay. "The Repository Jigsaw. "*Research Information*" May/June 2007.