PEER, Publishing and the Ecology of European Research, retour d'expérience de l'INRIA sur un projet européen associant éditeurs et archives ouvertes

Foudil Bretel, Alain Monteil

PEER, *Publishing and the Ecology of European Research*, retour d'expérience d'INRIA sur un projet européen associant éditeurs et archives ouvertes

Foudil Brétel[1] et Alain Monteil[2]

[1] Ingénieur DSI
INRIA Lyon la Doua, Bat. CEI-1, BP 52132, 66 bd. Neils Bohr
69100 Villeurbanne Cedex, France

[2] Ingénieur IST
INRIA Sophia Antipolis Méditerranée, BP 93, 2004 route des Lucioles
06902 Sophia Antipolis cedex, France

Résumé

L'INRIA (Institut national de recherche en informatique et automatique) participe au projet européen PEER, *Publishing and the Ecology of European Research*, qui regroupe éditeurs et représentants de la communauté scientifique. Nous présenterons ce projet qui a pour objectifs d'étudier l'appropriation et l'utilisation des Archives Ouvertes par les scientifiques, et de poser les fondements d'un nouveau système économique respectant à la fois les règles du marché de l'édition et le principe du libre accès à la connaissance. Puis nous détaillerons les avancées techniques développées autour de l'application servant d'interface entre les éditeurs et les archives ouvertes : le *PEER Depot*. Enfin nous évoquerons le bilan pour l'INRIA d'une participation à ce type de projet.

1. Introduction

Le monde éditorial suit de près les évolutions du paysage de la diffusion de la production scientifique, et notamment celui de la communication scientifique directe par les chercheurs vers les chercheurs. Plusieurs rapports et études ont été menés sur les serveurs de publications et la mise en ligne des documents, mais plus du point de vue des institutions ou communautés scientifiques que de celui de l'édition traditionnelle.

STM (*International Association of Scientific, Technical & Medical Publishers*) qui est l'association des plus grands éditeurs scientifiques a souhaité mener ce type d'étude afin d'avoir une vision claire de la situation. Le programme *eContent+* de l'Union européenne a permis de donner un cadre à une étude de cette nature.

2. Le projet PEER

2.1 L'avant-projet

Lors de la phase préliminaire au projet, STM s'est rapproché de la MPDL (*Max Planck Digital Library*) comme interlocuteur car la Max Planck est à la fois un institut de recherche pluridisciplinaire et un acteur majeur dans le libre accès. Laurent Romary qui était à l'époque directeur de la MPDL a aussitôt mentionné le cas français et souligné l'opportunité de faire participer une archive comme HAL (Hyper article en ligne, Centre pour la communication scientifique directe) puisque très bien positionnée dans le paysage scientifique hexagonale, et plus particulièrement l'INRIA comme partenaire technique.

STM a donc regroupé autour du projet une agence de financement, ESF (*European Science Foundation*), trois instituts de recherche UGOE (*Max Planck Digital Library*), MPG (*Max-Planck-Gesellschaft*) et INRIA afin de répondre à l'appel à projet *eContent+* dans le cadre du programme cadre *FP6* de l'Union européenne. Ce programme a pour objectif de favoriser et aider des initiatives de collaboration entre le monde du privé et les instituts de recherches publiques et mettre à disposition des citoyens du contenu scientifique. STM a fait appel à une consultante indépendante, durant six mois, pour l'animation et la récolte des informations nécessaires à la rédaction de la réponse de l'appel à projet.

Les compétences de cette assistance ont permis de compenser l'inexpérience des personnels IST en matière de projet européen. Nous pensons que c'est un des facteurs de réussite qui a permis au projet PEER de proposer un dossier solide et de qualité, et d'être retenu et financé par l'Union européenne.

2.2 Principes généraux

Le projet PEER propose la mise en place d'un observatoire, détaillé ci-après, à disposition des groupes de recherche afin d'étudier l'impact du libre accès dans l'écologie de l'édition scientifique européenne. Cet observatoire, après discussions, a été conçu comme un espace d'étude qui ne devait pas être « pollué » par des éléments extérieurs afin de refléter la réalité au plus près. C'est le premier principe de fonctionnement. Pour l'illustrer il a été décidé qu'aucune action de type incitation ou publicité ne serait faite par les archives participantes ou des documentalistes auprès des auteurs afin qu'ils déposent leurs articles. Pour cette raison, mais aussi pour des besoins techniques, chaque archive a bâti une archive parallèle et indépendante, telle l'instance PEER de HAL. Cette indépendance garantit des influences possibles vis-à-vis de l'archive institutionnelle.

Second principe général : celui de la définition des articles éligibles au projet PEER. STM a établi une liste de 240 titres de journaux représentant un échantillon relativement complet des différentes disciplines scientifiques, intégrant la notoriété par leur facteur d'impact et leur ancienneté. À partir de ce corpus, tous les articles dont l'auteur principal est résident d'un pays européen est éligible. Puis le volume est divisé en deux, une moitié sera diffusée par les éditeurs, l'autre par les auteurs.

Le dernier point fondamental du projet PEER concerne la définition du document devant être déposé. Il ne s'agissait pas de proposer les PDF des éditeurs, mais la version soumise et acceptée par le comité éditorial de la revue avant la mise en forme éditoriale. Ces

documents sont appelés « *Stage 2* » dans le cadre du projet PEER. Le « *Stage 1* » est la version auteur avant soumission et pour le « *Stage 3* » il s'agit de la version finale publiée par les éditeurs.

2.3 L'observatoire

L'observatoire est donc l'ensemble des processus et applications permettant au corpus d'articles scientifiques, rédigés par des scientifiques européens, d'être diffusé au travers des archives ouvertes. Le *PEER Depot*, réalisation principale de l'INRIA dans le projet, est au cœur du système et fait l'objet des avancées techniques les plus importantes. Elles sont détaillées dans la deuxième partie de cette contribution. Comme indiqué figure 1: l'alimentation devait se partager à égalité entre les éditeurs et les auteurs. Cette double alimentation doit être mesurable et observable au travers de logs afin d'étudier les usages concernant le dépôt dans cette archive ouverte.

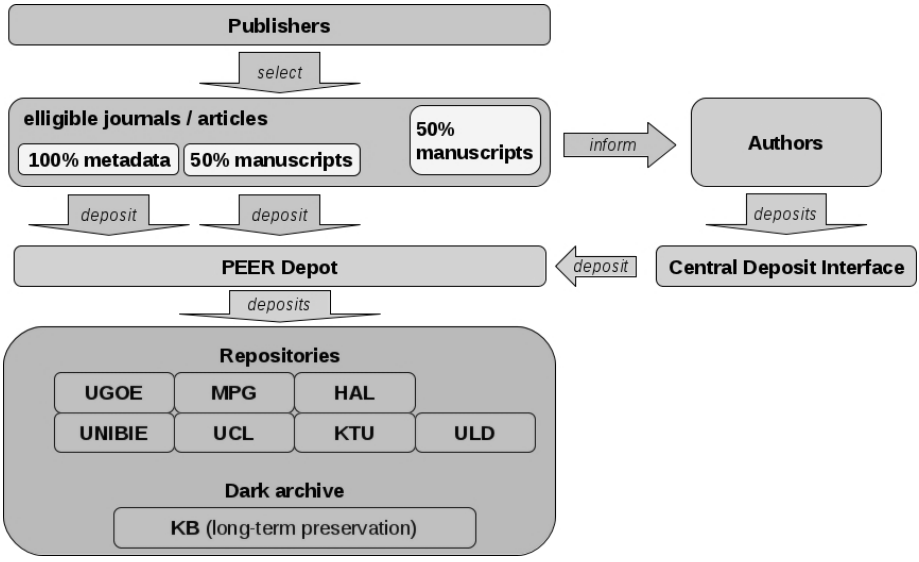

Figure 1 : Schéma des flux

Afin d'avoir matière à étudier, une masse critique de 10 000 références a été fixée par l'observatoire PEER. Au-delà de la volumétrie, chaque archive a également eu à mettre en place les éléments de recueil statistique, sans avoir de réel cahier des charges car les utilisateurs de ces données étaient inconnus.

L'atteinte de la masse critique n'a pas été aussi simple qu'il n'y paraît. Le temps de mise en place a été plus long que prévu, ce qui a entrainé un certain retard du projet. Pour ouvrir l'observatoire aux groupes de recherche, la masse critique devait être atteinte lors de la deuxième année. Cumulé au retard lié à la mise en place de l'observatoire, le dépôt direct par les auteurs n'a malheureusement pas été à la hauteur de nos espérances. Plus de 6 000 invitations aux dépôts ont été envoyées aux auteurs et seulement 147 ont répondu

favorablement par un dépôt d'article sur l'interface dédiée au projet PEER, gérée par la *Max Planck Digitial Library*.

Face à ce constat le projet PEER adopta des mesures pour atteindre l'objectif fixé. Première mesure : l'extension de 40 titres de journaux passant par le canal de dépôt éditeurs, les embargos ont été réduits accélérant la diffusion des articles, enfin ceux plus anciens ont été intégrés aux projets par certains éditeurs. Grâce à ces mesures le seuil a été atteint en mars 2011.

Concomitamment une demande de prolongation du projet a été demandée à l'Union européenne afin de donner suffisamment de temps aux groupes de recherche pour mener à bien leurs études. Prolongation accordée pour neuf mois supplémentaires portant à juin 2012 la fin du projet.

2.4 Objectifs

L'observatoire PEER doit permettre de déterminer les impacts du dépôt à grande échelle d'articles en libres accès. Son fonctionnement s'appuie sur trois groupes de recherche. Un premier prend en charge la thématique de l'usage des articles scientifiques tant côté plateforme éditeur que côté archives participantes par l'étude des logs et l'appréciation de la diffusion et l'accès *a priori* plus grands des résultats scientifiques. Le second groupe de recherche ; a pour objectif de suivre les tendances, expliquer les modèles éditoriaux des auteurs et le comportement des utilisateurs dans le contexte du libre accès, ainsi qu'e les impacts possibles sur la conduite et l'exercice de la recherche elle-même. Enfin, le dernier groupe étudie les aspects économiques avec pour objectif d'analyser les effets sur l'économie de la communication savante du dépôt de grande ampleur, d'enquêter sur le coût de celui à grande échelle, y compris du point de vue de l'efficacité économique ou sur le coût du processus de dépôt.

L'observatoire PEER permet également : d'étudier les facteurs influents sur le dépôt en libre accès des documents scientifiques, les facteurs favorisants mais aussi les freins ; d'étudier et produire des modèles éditoriaux illustrant la possible coexistence entre l'édition traditionnelle et l'auto-archivage en libre accès.

Le projet PEER et la mise en place de l'observatoire participent de la reprise d'une relation de confiance et de compréhension mutuelle entre les éditeurs et le monde académique soutenant les archives ouvertes.

Aujourd'hui plus de 16 000 articles sont disponibles grâce au projet PEER, et sont diffusés par les archives participantes, cependant seulement 0,2 % proviennent du dépôt des auteurs après invitation des éditeurs.

3. Brève description technique du *PEER Depot*

Le *PEER Depot* est l'intégration, sur un serveur *Linux RHEL 5*, d'une collection de *librairies* et scripts *Perl*, associés à un serveur FTP[1] *Pure-FTPd*. Les statistiques en vue de l'administration de la plateforme, ou à destination de l'équipe de recherche sur les usages, sont recueillies dans une base de données *PostgreSQL*. Le traitement est déclenché directement par le serveur FTP à la réception de nouveaux articles. Les dernières tâches (complétion des métadonnées, gestion de l'embargo, distribution vers les archives) sont automatisées par *cron*.

La plateforme conçue et développée à l'INRIA est constituée de logiciels libres. Les spécifications de départ indiquaient une moyenne de 50 articles à traiter par jour. Mais le serveur a régulièrement été amené à en traiter plusieurs milliers. La conception a suivi le principe KISS (*keep it simple and stupid*). Grâce à cette simplicité, on peut imaginer d'agréger plusieurs instances du *PEER Depot* en grappe, afin de supporter une plus forte montée en charge (*scalability*).

Une précision quant à la fonction du *PEER Depot* : nommé « *depot* » historiquement, il doit pourtant être considéré comme un *hub*. C'est-à-dire un relai de distribution. Les fonctions de stockage sont minimales. L'accent a été principalement mis sur le traitement et la distribution des articles.

4. Élaborer et gérer les flux : difficultés et solutions

4.1 Avancées organisationnelles

Un des grands accomplissements du projet, qui a significativement contribué à son succès, est l'établissement de conventions. Celles-ci ont nécessité un effort considérable de la part des participants, les monopolisant fortement durant les six premiers mois du projet, en générant de nombreuses réunions (partenaires PEER et éditeurs notamment). Ces conventions ont porté essentiellement sur le *workflow* de dépôt des éditeurs, et la fourniture des métadonnées [WP3.1]. Ce travail essentiel pour le *workpackage* 7 est dédié à l'élaboration de modèles pour la cohabitation de l'édition traditionnelle avec l'auto-archivage.

Lors de l'élaboration de ces conventions, l'équipe de recherche en charge des études des usages n'était pas encore constituée. Les partenaires impliqués dans le *workpackage 2* ont donc dû anticiper les futurs besoins (aussi bien pour les données de statistique que pour la fourniture des logs par les archives). Par la suite, la nécessité d'une collaboration plus rapprochée entre l'équipe de recherche et celle du *PEER Depot* ne s'est pas fait ressentir. On peut donc penser que les besoins ont été suffisamment bien estimés et couverts.

1 *File Transfer Protocol*.

4.1.1 Le dépôt éditeur

Si l'on considère le nombre d'organisations participantes (12 éditeurs, 6 archives), et leur diversité, en termes de contenu et de fonctionnement, l'échelle du projet était déjà une gageure, pas seulement pour l'établissement de conventions, mais aussi pour le traitement des spécificités de chacune. Ainsi au départ, il a été envisagé une période d'essai pour le dépôt des éditeurs à l'issue de laquelle ils devaient tous entrer simultanément en phase de production. Le délai nécessaire à l'adaptation technique (dénommée « validation » au sein du projet), aussi bien du *PEER Depot* que des éditeurs, n'a pas permis de respecter cette date commune. Aussi, le projet a dû se résoudre à traiter les cas un-à-un, avec une entrée de chaque éditeur séquentiellement.

DublinCore-like name	Comment
Title*	Article Title
Creator*	Corresponding Author's name: Last Name, First Name
AuthorEmail	Corresponding Author's email address
Description	Abstract
Date*	Date of Publication
Identifier*	DOI or PublisherArticleId[2]
Coverage	Geographic location of the Contributing Author: ISO 3166
Journal	Journal Title
Affiliation	Multi-tier organisation list: Country, Organization, Laboratory
ISSN	These elements are not mandatory to electronic publication, and can be derived from CrossRef after DOI is provided, and may therefore not be provided by publishers. Possible use of CrossRef for DOI resolution
Volume	
Issue	
Page	
Type*	Default value = article. Mapped to *info:eu-repo/semantics/article, info:eu-repo/semantics/acceptedVersion*
Subject	Subject headings; Scientific classification (defaults to what is provided in the general STM Journal table)
Language	ISO 639-3 (defaults to "eng")
Embargo	Embargo Period (defaults to what is provided in the general STM Journal table)

Tableau 1 : Métadonnées requises

[2] Identifiant article propre à l'éditeur.

4.1.2 Les métadonnées

La liste des métadonnées nécessaires, qui pouvaient être effectivement fournies par les éditeurs ou d'autres sources, a été établie par recensement auprès des archives participantes. Globalement, le nombre de métadonnées nécessaires était peu élevé. Le projet aurait pu se limiter aux recommandations de DRIVER[3], c'est-à-dire à 4 métadonnées issues du *Dublin Core*. Cependant, HAL et la *Koninklijke Bibliotheek*[4] en requéraient une douzaine. Il a donc été convenu d'en distinguer trois catégories : obligatoires, requises, optionnelles. Dans la pratique, tous les éditeurs à l'exception d'un ont été en mesure de fournir les métadonnées obligatoires et requises, les optionnelles quant à elles pouvaient être obtenues par d'autres sources (CrossRef ou fichiers).

Le tableau 1 est extrait du rapport D3.1[1] détaille les métadonnées utilisées par le projet.

4.1.3 Les standards

Se conformant aux recommandations de DRIVER, le projet a utilisé au maximum les standards (normes ISO notamment). Le DOI[5] y tient aussi une place particulière car les éditeurs avaient à cœur qu'ilsoit utilisé. En toute rigueur, cet identifiant renvoie au document publié, même si le projet traite du *stage 2*. Le DOI a cependant été retenu comme composante de l'identifiant des articles, sous la forme : PEER_stage2_[URL_encoded_DOI]. Le DOI est encodé pour des raisons pratiques de noms de fichiers qui ne peuvent pas contenir le caractère « / » (barre oblique) sous Linux. Ce caractère a posé problème à au moins un serveur Web d'archive pour lequel une option de sécurité précisait que les requêtes HTTP[6] ne peuvent pas contenir de barre oblique. L'archive a décidé de renommer les fichiers en transformant « / » en « _ ». Ce qui n'a pas particulièrement perturbé l'équipe de recherche en charge des études des usages, qui fonde ses recherches sur les URL enregistrées dans les fichiers logs des serveurs Web des archives.

4.1.4 Le filtrage

Hormis la distribution des articles proprement dite, le rôle du *PEER Depot* devait recouvrir également des fonctions de filtrage selon plusieurs critères : période d'embargo, type d'article, contenu européen, éventuellement thème[7].

La gestion de l'embargo était une préoccupation importante pour les éditeurs. Le calcul de celui-ci est identique dans les deux cas de flux : dépôt auteur et éditeur. Dans la mesure où les éditeurs fournissaient une date de publication (papier ou en ligne), et où le *PEER Depot* disposait des durées d'embargo par journal, leur gestion n'a pas posé problème. Voir ci-dessous *Interface de dépôt auteur* pour plus de détails.

Ne pouvant les prendre à leur charge, beaucoup d'éditeurs ont souhaité que des filtrages de contenu soient effectués sur le *PEER Depot* : par type d'article ou par journal. Ces informations font partie des métadonnées requises par PEER. La principale difficulté a été

---

3 *Digital Repository Infrastructure Vision for European Resarch.*
4 Bibliothèque royale des Pays-Bas.
5 *Digital Object Identifier*.
6 *Hypertext Transfer Protocol*.
7 Cas de l'archive thématique *Social Science Open Access Repository* (SSOAR).

d'adapter le filtrage, tout au long du projet, aux spécificités variables des éditeurs (qui n'utilisent pas toujours de manière orthodoxe les champs dédiés).

Pour le filtrage du contenu européen, il a été convenu qu'il serait basé sur le pays de l'auteur correspondant. Cette donnée était issue des affiliations (métadonnées requises). Les noms de pays et codes divers étant transformés en ISO-3611-A2. La possibilité d'inclure le pays de l'auteur principal dans la sélection ainsi que celui des autres auteurs a été évoquée en cours de projet. Elle n'a pas été retenue car l'origine de chacun n'est pas systématiquement fournie par les éditeurs.

Le filtrage par thème a la particularité d'être effectué en sortie du *PEER Depot*, tandis que tous les autres filtrages s'effectuent à l'entrée. En effet, il ne concerne pas toutes les archives. Le critère est dans ce cas l'ISSN[8]. Le filtrage en sortie étant assez modulaire de conception, d'autres solutions et critères auraient pu être envisagés aisément.

4.1.5 La distribution vers les archives

4.1.5.1 Le format d'échange

Concernant l'interfaçage entre le *PEER Depot* et les archives, la principale question a été celle du format de livraison des articles, et particulièrement le choix du format XML[9]. Les formats en lice étaient NLM[10], format adopté par 6 des 12 éditeurs, et TEI[11], standard parfaitement maitrisé par l'équipe INRIA, et largement utilisé dans l'édition scientifique et le monde académique. Le choix s'est arrêté sur la TEI. Un schéma *RelaxNG* est en cours d'élaboration afin de décrire formellement le format TEI utilisé dans PEER.

Ce choix de la TEI comme format pivot unique s'inscrit dans une démarche de standardisation. En effet, il s'agit d'une initiative internationale de référence pour les sciences humaines numériques, utilisée par les principaux acteurs institutionnels en édition électronique française (Persée, Revues.org, Presses universitaires de Caen). Elle propose une plateforme unifiée de description de documents textuels (manuscrits, livres, dictionnaires... et articles), ce qui permet d'éviter l'effet de silo et la fragmentation des formats créés par la NLM. Le *framework* est donc modulaire, extensible et propose plus de 500 éléments, y compris des extensions telles que SVG[12], MathML CALS, etc. Elle s'accompagne d'outils précieux, tels qu'un langage de spécification (ODD[13]) permettant d'adapter facilement un schéma à ses besoins et de le générer automatiquement (DTD[14], XML ou RelaxNG[15]). Enfin, elle est gérée de manière consensuelle et ouverte, par une communauté très active[16].

---

8 *International Standard Serial Number*.
9 *Extensible Markup Language*.
10 *National Library of Medicine.*
11 *Text Encoding Initiative.*
12 *Scalable Vector Graphics*.
13 One Document Does it all.
14 *Document Type Definition*.
15 *REgular LAnguage for XML Next Generation*.
16 Pour une réflexion sur les formats XML, voir [2].

4.1.5.2 Le protocole de dépôt

Le projet a choisi d'utiliser le protocole SWORD[17] pour la distribution des articles du *PEER Depot* vers les archives. Toutes ont eu à implémenter le support de SWORD. Dspace[18] incluant nativement ce support, les archives qui utilisaient ce logiciel n'ont eu qu'à adapter la transformation (XSLT[19]) de TEI vers leur propre format de métadonnées. L'INRIA ayant fourni un modèle générique. SWORD avait déjà été adopté par quelques rares archives (ArXiv par exemple), mais l'utilisation d'un protocole standard unique était une nouveauté. D'ailleurs SWORD étant assez souple (il continue d'évoluer), cependant l'implémentation n'a pas été comprise de la même manière par tous. L'implémentation retenue dans PEER ne convenait pas exactement aux archives pour lesquelles le dépôt s'effectuait en plusieurs étapes (*via* une validation humaine par exemple). L'avantage principal de SWORD pour le projet est le mécanisme d'acquittement qu'il inclut : l'archive transmet directement dans sa réponse l'identifiant du document déposé.

4.1.6 Interface de dépôt auteur

La spécification initiale du dépôt auteur n'avait pas prévu que, hormis HAL, les archives étant des archives institutionnelles, elles n'autoriseraient le dépôt qu'aux utilisateurs de leur propre institution. Il a donc été nécessaire de concevoir une interface de dépôt auteur générique. Hébergée à la *Max Planck Digital Library* (MPDL), elle transmet ces dépôts au *PEER Depot*, qui les distribue aux archives à la fin de la période d'embargo (de 0 à 36 mois). Cette date de distribution est calculée selon la formule : date de publication + période d'embargo = date de distribution. La date de publication est issue des métadonnées fournies par les éditeurs. La durée d'embargo définie par journal est issue de la liste des journaux intégrée au *PEER Depot*. Les éditeurs fournissent les métadonnées pour les journaux à destination du flux « éditeur », mais également celles à destination du flux « auteur ». Le *PEER Depot* se charge d'établir la corrélation, à partir d'un petit ensemble de métadonnées fournies par l'auteur. Dans la pratique, l'appariement s'effectue sur le titre du journal et celui de l'article uniquement[20].

4.2 Avancées techniques

4.2.1 Le développement

La conception et les spécifications du *workflow* de dépôt des éditeurs ayant requis plus de temps que prévu, le développement lui s'est réalisé dans des délais très courts, et a représenté environ 3 mois.homme pour le premier prototype. Ce qui comprend : l'écriture des *XSLT*, du code *Perl*, l'intégration de la plateforme, et la configuration système. Les besoins

---

17 *"a JISC-funded initiative to define and develop a standard mechanism for depositing into repositories*". SWORD est un profile du protocole *Atom Publishing Protocol* (APP). *Cf*. <http://swordapp.org/> consulté le 22 mai 2012.

18 Logiciel libre utile à la construction d'archives électroniques ouvertes.

19 *Extensible Stylesheet Language Transformations*.

20 Voir le projet similaire *Open Access-Repository Junction* : <http://oarepojunction.wordpress.com/about-the-broker/> consulté le 22 mai 2012.

étant très spécifiques, pratiquement aucun code existant n'a pu être réutilisé, hormis quelques *librairies* fondamentales telles que XML::LibXML. Une tentative d'associer d'autres partenaires du projet au développement n'a malheureusement pas abouti, principalement pour des raisons d'usages techniques différents.

4.2.2 Fusion XML

Pour chaque article les métadonnées pouvaient être fournies en deux passes. Le seul vrai défi technique a été la fusion XML des métadonnées. Elle survient à la seconde passe. Par convention, l'opération se limite à la complétion des métadonnées, ce qui simplifie le problème. C'est-à-dire qu'il n'y a pas d'écrasement : seules les données dont on ne dispose pas déjà sont prises en compte. Cependant, comment comparer des nœuds pour lesquels l'information discriminante se situe plusieurs niveaux en dessous ? L'exemple ci-dessous illustre ce problème :

```
<author>
  <persName>
    <forename type="first">Foudil</forename>
    <surname>Brétel</surname>
  </persName>
  <email>foudil.bretel@INRIA.fr</email>
  <affiliation>
    <orgName type="department">Direction des Systèmes d'Information</orgName>
    <orgName type="institution">INRIA</orgName>
    <address>
      <postCode>75001</postCode>
      <settlement>Paris</settlement>
      <country key="FR">France</country>
    </address>
  </affiliation>
</author>
```

Dans cet exemple, lorsqu'on considère le nœud *author*, il faut prendre en compte ses descendants afin de pouvoir l'identifier (en vue de le comparer et de l'ajouter éventuellement). Sans entrer dans les détails, nous avons adopté un algorithme original, qui compare les descendants sur au plus deux niveaux. Celui-ci a donné des résultats satisfaisants.

4.2.3 Tests XML

Nous n'avons pas pu implémenter suffisamment de tests pour couvrir toute l'application – notamment en raison de la difficulté à automatiser les tests d'un *workflow* complet avec un espace de stockage de test spécifique. En revanche, il est devenu urgent de se prémunir d'effets de bords indésirables lors de modifications des transformations XSLT. Puisqu'il n'existait pas vraiment d'outils simples dédiés, nous avons élaboré le nôtre. Nous disposons d'un jeu d'environs 150 tests. Notre réalisation a montré des performances supérieures aux dispositifs similaires.

4.2.4 Conversion des fichiers source

En première analyse, et sur demande des éditeurs, notre équipe avait proposé la possibilité de convertir, sur le *PEER Depot*, lorsque le fichier PDF n'était pas disponible, les fichiers sources des articles, du format MS Word ou LaTex vers PDF. En dernier recours, cette technologie aurait pu être empruntée au CCSD[21]. Mais les échantillons ont montré que cela n'était pas réalisable du fait de la diversité des sources : le fichier MS Word ne contenait pas toujours les figures, qui pouvaient être placées dans des sous-répertoires variés. La conversion n'était donc pas automatisable, en tout cas pas sans une convention préalable. Le *PEER Depot* a bien reçu des fichiers sources sans PDF, et dans ce cas, les articles ont été considérés comme perdus pour le projet.

4.2.5 Confidentialité

La question de la confidentialité des fichiers de log fournis par les archives à l'équipe de recherche sur les usages s'est posée tôt, et de manière plus insistante pour HAL et la MPDL, qui obéissent à des législations plus restrictives dans ce domaine (comparées par exemple à celle du Royaume-Uni). La solution a été laissée à la discrétion de chaque archive, tant que les origines des requêtes HTTP pouvaient être identifiées. Ainsi pour HAL, les logs sont traités avant livraison afin de transformer l'adresse *IP* et le champ *Referrer* en des identifiants uniques (SHA1).

4.2.6 Reporting

Le *reporting* sur l'activité du *PEER Depot* a aussi été envisagé très tôt. Hormis l'équipe de recherche sur les usages, d'autres instances du projet devaient pouvoir accéder aux données enregistrées par le *PEER Depot*. Le choix de l'outil de stockage s'est donc naturellement porté sur une base de donnée relationnelle (*PostgreSQL*), dont l'accès a été ouvert aux personnes intéressées. Il n'a pas été nécessaire de développer une interface spécifique pour la consultation de l'activité : la structure de la base de données restant simple, les outils de consultation de bases SQL (tel *pgAdmin*) ont suffi. En revanche, l'équipe en charge du monitoring a élaboré un mécanisme de génération automatique de rapports exhaustifs.

4.2.7 GROBIB

Le projet a été une nouvelle occasion d'expérimenter GROBID, l'automate d'extraction de données bibliographiques à partir de documents PDF, développé en collaboration avec

21 Centre pour la communication scientifique directe.

l'INRIA [3]. Le besoin est apparu avec un éditeur qui ne pouvait pas facilement fournir les affiliations. La solution adoptée a été d'extraire l'ensemble des métadonnées des articles en PDF, et de les compléter éventuellement par d'autres sources ; celles fournies par l'éditeur étant ignorées. Cette solution a nécessité des ajustements, en termes de format (se conformer à la TEI telle qu'utilisée dans PEER) et de contenu.

Les résultats sont encourageants (environ 2 200 articles traités, dont 950 retenus), et la solution, dans le cadre de PEER, pourrait être encore améliorée : surtout en automatisant le traitement, qui souffre d'être actuellement semi manuel, avec un *workflow* spécifique parallèle, sur un mode « essai-erreur », et donc enclin aux erreurs.

4.3 Contingences

Des difficultés sont survenues en cours de projet. En partie parce qu'elles n'avaient pas été abordées dès la conception, ou pas suffisamment développées, en partie parce que l'état de l'art ne permettait pas de les pallier.

4.3.1 Constat sur qualité des données

Une première surprise a concerné la qualité des données. Celles dont disposent les éditeurs proviennent souvent directement des auteurs, qui n'ont pas les moyens de les consolider, comme en les extrayant de référentiels. On pense notamment aux affiliations, mais on pourrait aussi évoquer les codes pays... Une des conséquences est qu'une partie non négligeable des articles restent incomplets en termes de métadonnées, et donc en attente sur le *PEER Depot*.

En termes techniques, les archives, qui étaient associées à l'élaboration du *workflow*, ont été assez autonomes. Cependant, il est intéressant de constater que toutes ont déployé une archive distincte dédiée à PEER. Les raisons sont d'ordre éditorial pour les archives de type institutionnel (PEER délivre des articles sans distinction d'affiliation d'auteur, or les archives souhaitent recevoir les articles de leur organisation), et d'ordre technique pour la seule archive de type national (les contraintes sur les métadonnées ne pouvant pas être satisfaites). On peut déplorer que la communauté de l'IST n'ait pas encore réussi à s'organiser afin de proposer un référentiel unique des affiliations.

4.3.2 Visibilité des articles déposés

Les équipes de recherche du projet (usages et comportements principalement) étaient préoccupées par la visibilité des articles déposés dans ces instances d'archives ouvertes dédiées à PEER. Cette visibilité était souhaitée identique à celle des instances principales. Les équipes de recherche prévoyaient notamment que les chercheurs-lecteurs parviendraient à ces instances par l'intermédiaire des moteurs de recherche (tel *GoogleScholar* notamment). Or la visibilité dans ces moteurs n'est pas maitrisable : on ne peut que déclarer son site au moyen d'un *sitemap*[22], puis attendre que Google indexe les notices. Il est à noter que *GoogleScholar* ne supporte plus OAI-PMH pour l'indexation, mais *OpenURL*. Cela a exclu une archive candidate[23] du projet : par manque de ressources pour l'implémentation.

[22] <http://www.sitemaps.org/> consulté le 22 mai 2012.

[23] *Kaunas University of Technology*.

Globalement, une certaine visibilité a été atteinte, mais avec des résultats disparates d'une archive à l'autre.

4.3.3 Quantité de contenu déposé

À un certain stade du projet, l'équipe de recherche sur les usages a considéré que la masse critique d'articles déposés nécessaire à la recherche ne serait pas atteinte en temps voulu. Le projet a donc été prolongé de 9 mois. Et la quantité de contenu déposé dans les archives est devenue une priorité. Le projet a pu atteindre son nouvel objectif (10 000 articles déposés fin février 2011, soit un an après les premiers dépôts), en prenant des mesures dont certaines modifiaient sensiblement les conditions initiales, dont :

- passage de certains journaux du flux auteur vers le flux éditeurs ;
- inclusion de certains types d'articles ;
- réduction d'embargos pour certains journaux ;
- inclusion de collections plus anciennes.

4.3.4 Retirer des articles de PEER

La possibilité d'un retrait d'article à la demande d'un auteur avait été envisagée assez tôt, et les archives s'étaient engagées à permettre une forme de retrait. Le détail de ce processus sur l'ensemble du *workflow* n'a pas été spécifié et n'a été implémenté que lorsque le besoin s'est déclaré. Or celui-ci est venu des éditeurs, et sur des collections entières (ce qui a représenté parfois des centaines d'articles). Les principales raisons ont été : la présence de commentaires des relecteurs ou des auteurs, et du contenu manquant. La difficulté était de garder une cohérence sur l'ensemble du *workflow*. La modalité retenue a été la suivante : demande de retrait d'un auteur ou d'un éditeur auprès des interlocuteurs désignés sous forme d'une liste de DOI ; le retrait est effectué sur le *PEER Depot* qui en garde la mémoire (et donc de toutes les informations associées dont les dépôts correspondants vers les archives) ; le *PEER Depot* informe les archives du retrait au moyen d'une liste de DOI avec les identifiants propres à chaque archive ; les interlocuteurs contrôlent le retrait effectif des archives.

5. Bilan et perspectives

5.1 En termes métier

Pour le réseau IST et pour INRIA la participation au projet PEER fut une première dont les principaux enseignements peuvent être tirés.

En interne, en termes de support administratif nous avons rencontré des difficultés d'accompagnement. En effet la légitimité pour les services administratifs d'accompagner une ligne métier et non des scientifiques pour un projet européen n'était pas évidente. Le manque d'effectifs et le nombre important de projets européens obtenus par INRIA nous a parfois fait passer à un niveau inférieur de priorité. Nous avons cependant pu bénéficier de ce soutien indispensable pour la bonne marche d'un tel projet.

L'expérience du réseau IST en matière de gestion de projets européens était très faible, et principalement réduite à celles du conseiller à l'IST Laurent Romary, du délégué à l'IST Jacques Millet, et Alain Monteil chef de projet HAL-INRIA relativement novice ! La participation a été variable dans le temps, assez forte au début du projet puis de plus en plus faible, sauf pour J. Millet membre de l'*executive board*[24]. Cette expérience nous a permis d'accroître notre connaissance du paysage des archives ouvertes en Europe, ainsi que de la vision que porte le monde éditorial sur le libre accès.

La confrontation avec nos partenaires d'autres structures et d'autres pays a permis de partager des expériences, de mieux se connaître permettant d'initier de possibles partenariats. Nous avons effectué une visite auprès de l'équipe de la MPDL à Munich afin de partager nos pratiques en matière d'archives ouvertes. Autres points positifs à signaler, le dialogue et contact direct avec les éditeurs autour des échanges d'information et le traitement des documents pour une bonne gestion et signalement dans les archives ouvertes. L'expertise INRIA en matière IST est à présent mieux connue et nous pouvons nous mettre en ordre de marche pour d'autres projets européens comme partenaire potentiel. Il est à signaler que la France est peu représentée dans ce type d'action, nous devrions être plus pro-actif en la matière à l'avenir.

5.2 En termes technique

La réalisation et l'exploitation du *PEER Depot* est un des éléments forts de réussite du projet PEER et de visibilité des compétences d'INRIA. Le traitement des données comme la mise au format TEI ou l'extraction de métadonnées ont permis à ses ingénieurs d'acquérir une haute expertise reconnue. Les avancées devront être diffusées et réutilisables par tous, notamment l'extraction de métadonnées à partir des fichiers sources qui pourra être intégrée comme une offre de service d'assistance au dépôt dans les archives ouvertes.

Côté éditeurs, les solutions techniques permettent de diffuser *via* des archives ouvertes institutionnelles la documentation scientifique, par ailleurs elles ouvrent des opportunités pour le modèle auteur-payeur. L'exemple de BioMedCentral et la diffusion *via* HAL-INSERM en utilisant le protocole SWORD, également utilisé dans le cadre du projet PEER, est emblématique de ce que pourrait être la communication scientifique ouverte.

En matière de perspectives, le projet PEER apportera à la fin un corpus de plusieurs milliers d'articles scientifiques librement accessibles, conservés de manière pérenne. Chaque archive participante pourra si elle le souhaite intégrer ce corpus dans sa base.

Participer au projet PEER a apporté une sensibilité nouvelle au réseau IST, visible par l'intérêt pour d'autres projets comme OpenAIRE et MedOANet. Cette expérience du réseau IST INRIA pourra être capitalisée en vue de la participation à d'autres projets de ce type.

Bibliographie

[1] The eContent*plus* Programme <http://ec.europa.eu/information_society/activities/econtentplus/closedcalls/econtentplus/programme/index_en.htm> consulté le 22 mai 2012.

[24] Organe de pilotage du projet réunissant principalement les représentants des partenaires.

[2] Les rapports du projet PEER : <http://www.peerproject.eu/reports/> consulté le 22 mai 2012.

[3] Publications et présentations autour du projet PEER : <http://www.peerproject.eu/publications-presentations/> consulté le 22 mai 2012.

[4] M. Bijsterbosch, F. Brétel, N. Bulatovic, D. Peters, M. Vanderfeesten, J. Wallace, "Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving", May 2009. <http://www.peerproject.eu/reports/> consulté le 22 mai 2012.

[5] J. McDonough, "Structural Metadata and the Social Limitation of Interoperability: A Sociotechnical View of XML and Digital Library Standards Development," communication à The Balisage: The Markup Conference, Montréal, Canada (2008).

[6] P. Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," dans M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas, Research and Advanced Technology for Digital Libraries. Éd. Berlin, Heidelberg, 5714, p. 473-474 (2009).

[7] L. Romary, "TEI and Scholarly publishing – experience from the PEER project", <http://dho.ie/sites/default/files/events/teipublishing/TEIScholPub_Dublin.pdf>. consulté le 22 mai 2012.

[8] Schematic of PEER Depot Workflow after *PEER* Annual Report D9.8. ECP-2007-DILI-537003. PEER. Annual Report – Year 2. (2010) <http://www.peerproject.eu/fileadmin/media/reports/D9_8_annual_public_report_20100930.pdf>. consulté le 22 mai 2012.

[9] F. Bretel, P. Lopez, M. Medves, A. Monteil, L. Romary, "Back to meaning – information structuring in the PEER project". Author manuscript, published in TEI Conference (2010) <http://www.peerproject.eu/fileadmin/media/ppt_about_peer/PEERBreakingNews.pdf> consulté le 22 mai 2012.

[10] P. Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications". In Proceedings of ECDL, 13th European Conference on Digital Library, Corfu, Greece (2009).

[11] J. Fry, C. Oppenheim, S. Probets, C. Creaser, H. Greenwood, V. Spezi, S. White, "D4.1 PEER Behavioural Research: Authors and Users vis-à-vis Journals and Repositories – Baseline report" <http://www.peerproject.eu/fileadmin/media/reports/Final_revision_-_behavioural_baseline_report_-_20_01_10.pdf> consulté le 22 mai 2012.

[12] P. Dubini, "Complementary Article Dissemination via Journals and Repositories: Economic Evidence from the PEER Project", presented at APE 2011 - Academic Publishing in Europe, Berlin Brandenburg Academy of Sciences, Berlin, 11-12 January 2011. <http://www.peerproject.eu/fileadmin/media/ppt_about_peer/APE_presentation.pdf> consulté le 22 mai 2012.