# Les statistiques d'utilisation d'archives ouvertes - État de l'art

Joachim Schöpfel, Hélène Prost

▶ **To cite this version:**

Joachim Schöpfel, Hélène Prost. Les statistiques d'utilisation d'archives ouvertes - État de l'art. 2010. sic_00480538

**HAL Id: sic_00480538**

**https://archivesic.ccsd.cnrs.fr/sic_00480538**

Preprint submitted on 4 May 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Les statistiques d'utilisation d'archives ouvertes - Etat de l'art

**Joachim Schöpfel** est maître de conférences en sciences de l'information et de la communication et directeur de l'UFR IDIST de l'université Charles-de-Gaulle Lille 3.

Université Charles-de-Gaulle Lille 3, UFR IDIST, BP 60149, 59653 Villeneuve d'Ascq Cedex.

joachim.schopfel@univ-lille3.fr

**Hélène Prost** est chargée d'étude à l'INIST-CNRS.

INIST-CNRS, 2 allée du Parc de Brabois, 54519 Vandoeuvre-lès-Nancy Cedex.

helene.prost@inist.fr

## *Résumé*

Cette communication présente les premiers résultats d'une étude sur les archives ouvertes en France à partir d'informations trouvées dans des publications ou sur les sites web propres aux archives. L'objectif de cette étude est de faire un bilan des résultats de la politique en faveur des archives ouvertes en France en termes de typologie, contenu, taille, développement et utilisation. Dans un 2e temps nous ferons l'état de l'art de la question des statistiques d'utilisation des archives ouvertes, avec quelques recommandations pour leur normalisation.

## *Mots-clés*

Archives ouvertes, accès libre, statistiques d'utilisation, information scientifique

## *1. Introduction*

Dans l'environnement du mouvement vers l'accès libre à l'information scientifique, les revues gratuites en ligne et les archives ouvertes sont devenues en quelques années une partie significative du paysage de la recherche, estimée à 15-20% de la production scientifique (cf. [JANIK, AUBRY, 2005], [WILLINSKY 2006] ou [LUTZ 2009]).

Institutions, organisations et gouvernements investissent dans la mise en place, l'infrastructure, la gestion et la maintenance de ces nouveaux outils, car l'accès libre à l'information scientifique, la communication rapide, directe et non restrictive entre chercheurs n'est pas seulement utile à la recherche mais est devenu un enjeu politique d'envergure. La Commission Européenne soutient le développement d'une infrastructure

qui facilite la libre circulation de l'information, la 4e liberté de l'espace de l'Union et condition indispensable de la société de l'information.

Néanmoins les études empiriques sur l'impact réel de ce mouvement, en termes de développement et surtout d'utilisation, sont plutôt rares. Pourtant, dans le domaine des archives ouvertes les enjeux ne manquent pas et les statistiques d'utilisation intéressent tous les acteurs concernés [CARR et al., 2008].

En 2008, nous avions mené une première analyse de la réalité sur le terrain [SCHÖPFEL, STOCK, 2009a, b]. D'après nos chiffres, le nombre des archives ouvertes en France s'élevait en 2008 (mars-mai) à 56 dont 48% relevaient de l'enseignement supérieur. Le nombre total des dépôts dépassait 700 000 documents, données (observations, résultats d'analyses), notices etc.

La 2e phase de notre étude poursuit un double objectif : continuer l'analyse du développement des archives ouvertes en France (follow-up) et faire le lien avec les autres recherches de notre équipe à Lille sur les usages des ressources numériques en ligne.

## *2. Méthodologie*

Notre analyse s'appuie sur une adaptation de la méthodologie développée pour l'étude en 2008. Comme en 2008, nous avons sélectionné les archives ouvertes à partir d'un choix de sites Web de référencement (répertoires), au lieu de procéder à une recherche directe sur le Web, ou de mener une enquête auprès des institutions. L'analyse porte donc uniquement sur des sites référencés et répertoriés, validés par d'autres comités professionnels ou scientifiques, avec un rayonnement et une visibilité nationale et/ou internationale.

Aux répertoires de 2008[1], nous avons ajouté onze autres sites dont le wiki « Archives Ouvertes » des URFIST, la liste des archives ouvertes de la Mission IST du MESR, le site du Groupe de Travail Archives Ouvertes du consortium COUPERIN et la liste sur le site du CCSD.

En recoupant ces répertoires, nous avons établi une liste d'archives ouvertes. Pour chaque entrée nous avons vérifié l'URL, la localisation en France et la présence de dépôts récents, en écartant les doublons.

Chacune des archives a été caractérisée selon les 58 critères d'une grille d'analyse. Les résultats ont été intégrés dans une base de données, vérifiés, validés et le cas échéant, modifiés et/ou complétés par une ou deux personnes avant l'analyse statistique ou qualitative. Cette base de données contient avec 8 700 entrées quatre fois plus d'information que celle de 2008 (1 960 entrées). Trois types d'analyse sont effectués : une analyse statistique des données chiffrées, une analyse comparative entre les données 2008 et 2009, et une analyse qualitative. L'analyse se poursuit à l'heure de la rédaction de cette communication.

Pour toutes les archives, nous avons cherché des données d'usage (accès en ligne). En 2008, les sites des archives ouvertes ne mettaient que peu d'éléments chiffrés en ligne, et la

[1] Parmi ces répertoires figuraient notamment Eprints, OpenDOAR et ROAR.

publication des statistiques restait l'exception.

Le planning de l'étude s'est déroulé suivant ces différentes étapes :

Choix des sites de référencement : mai 2009

Sélection des archives ouvertes : mai-juin 2009

Caractérisation des archives ouvertes : juillet-octobre 2009

Recensement des données d'utilisation : octobre-novembre 2009

L'état de l'art a été réalisé en parallèle.

L'étude de cas du site IRIS[2] a commencé en octobre 2009 mais se poursuivra jusqu'en février ou mars 2010.

## 3. Résultats

Nous présentons ici l'évolution du paysage des archives ouvertes en France entre 2008 et 2009 et quelques résultats de leur utilisation.

### 3.1. Développement des archives en termes d'offre

La comparaison des résultats entre 2008 et 2009 montre un spectaculaire accroissement de l'offre, tant qu'en nombre d'archives qu'en nombre de documents. En 2008, 56 archives sont recensées; en 2009, s'y ajoutent 94 autres archives. On dénombre 703 178 items en 2008 ; puis 1 878 520 en 2009.

#### 3.1.1. Institutions et typologie des archives

La plupart des différentes institutions qui ont en charge la gestion des archives appartiennent soit à la recherche publique, soit à l'enseignement supérieur. Entre 2008 et 2009, on note un engagement plus fort de la recherche publique qui gère un peu plus d'archives (67 contre 60 gérées par l'enseignement supérieur) et signale un nombre plus important de documents (56% des documents signalés contre 28% par l'enseignement supérieur).

Une archive se définit selon sa typologie [ARMBRUSTER, ROMARY, 2009]. Une archive est institutionnelle si elle regroupe les différentes publications d'une même institution ; au nombre de 32 en 2008, puis 87 en 2009, ce type est le plus représenté dans l'échantillon de 2009.

Une archive est thématique (*subject-based*) si sa caractéristique principale est de regrouper des documents se rapportant à un sujet commun ; nous avons identifié 13 et 20, respectivement en 2008 et 2009.

Finalement, certaines archives contiennent un seul type de document, souvent des thèses. Au nombre de 4 en 2008, on en compte 20 en 2009.

[2] Archive institutionnelle de l'université Lille 1 (USTL).

Une archive peut à la fois être institutionnelle et ne contenir que les thèses soutenues dans cette institution.

Nous avons également défini les archives selon la politique de dépôt : les trois principales sont, l'incitation pour 86 archives en 2009, l'obligation pour 16 archives ou la politique patrimoniale pour 30 sites. Deux archives sont régies par une politique mixte : patrimoniale et incitative pour IRIS, incitative ou obligatoire selon les établissements regroupés sous l'archive OATAO[3].

En ce qui concerne le signalement des documents, l'interface la plus répandue est le dépôt (103 sites) où le chercheur peut signaler ses publications de sa propre initiative ou aidé par les gestionnaires du site. Ce modèle reflète le principe de base d'une archive ouverte. Par contre, parmi les sites référencés comme « archives ouvertes » on dénombre aussi 15 bibliothèques numériques, 19 bases de données et 12 sites web.

### *3.1.2. Evolution des contenus*

Entre 2008 et 2009, le nombre d'items contenus dans les archives ouvertes augmente de 167%. Voici les 5 sites les plus importants en terme de dépôts en 2008 et en 2009 (tableau 1).

| | **Acronym** | **Name** | **Items 2008** |
|---|---|---|---|
| 1 | PERSEE | Revues Scientifiques en Sciences Humaines et Sociales | 172 215 |
| 2 | HAL | Hyperarticle en Ligne | 108 590 |
| 3 | ProdINRA | Base de données des publications de l'INRA | 100 000 (appr.) |
| 4 | COD | Cristallography Open Database | 70 295 |
| 5 | IRD | IRD Horizon Pleins Textes | 68 519 |

| | **Acronym** | **Name** | **Items 2009** |
|---|---|---|---|
| 1 | PERSEE | Revues Scientifiques en Sciences Humaines et Sociale | 259 816 |
| 2 | Gallica | Gallica | 215 422 |
| 3 | HAL | Hyperarticle en Ligne | 143 341 |
| 4 | ProdINRA | Base de données des publications de l'INRA | 118 543 |
| 5 | COD | Cristallography Open Database | 110 210 |

**Tableau 1 : Classement des 5 premiers sites en 2008 et 2009**.

[3] Archive institutionnelle de l'Institut National Polytechnique de Toulouse et de l'Ecole Vétérinaire de Toulouse

Ce classement a deux particularités. D'une part, la présence d'une archive de données (*datasets*) sans documents (COD), précurseur si on veut de la cyberinfrastructure de recherche. L'autre particularité est la présence de deux sites qui sont référencés comme archive ouverte mais qui n'ont pas la vocation d'une communication scientifique directe : Gallica de la BNF d'une part (avec surtout des documents et images numérisés tombés dans le domaine public), et PERSEE, l'archive numérique des revues en SHS.

Les sites contenant de la littérature grise sont répartis dans les mêmes proportions, 75% en 2008 et 74% en 2009 ; par contre, la part des documents gris augmente, passant de 11% en 2008 à 17% en 2009. En terme de chiffre brut, le nombre total de documents gris augmente de 301%, de 79 005 en 2008 à 316 751 en 2009. Alors que les thèses constituaient le type de littérature grise le plus important en 2008 (46%), ce sont les conférences qui arrivent en tête en 2009, en représentant 49% des documents gris.

16% des dépôts (items) dans les archives sont des notices bibliographiques, sans accès au texte intégral. Dans 37 archives, la part des notices sans texte intégral est égale ou supérieure à 50%.

### *3.2. Statistiques d'utilisation d'archives ouvertes*

Nous avons trouvé des informations sur les statistiques d'utilisation pour dix sites. C'est peu représentatif, comparé au nombre total (= 7%), mais c'est un progrès par rapport à 2008 où un seul site communiquait sur ses chiffres d'utilisation.

Ceci étant, certaines informations font un amalgame entre « utilisation-dépôt » et « utilisation-accès/téléchargement ». Ainsi, nous avons trouvé deux bilans où sous la rubrique « utilisation » seul le nombre des dépôts avait été compté [GOUAT, 2009] ; [TABORELLI, 2005].

Les autres bilans ou rapports sur l'utilisation des archives ouvertes donnent trois types d'information – une information globale sur l'accès ou le téléchargement des documents par auteur ou institution, une information par type de documents, et une analyse détaillée des fichiers log.

#### *3.2.1. Statistiques à destination des auteurs et institutions*

Toutes les archives de HAL fournissent des statistiques d'utilisation aux dépositaires (auteurs, institutions) et auteurs enregistrés. Ces statistiques sont fournies pour chaque document et pour l'ensemble des documents, avec le format suivant :

- La durée de la mise en ligne (en année, moi, jour).
- Accès à la fiche concise des métadonnées.
- Accès à la fiche étendue (= détaillée) des métadonnées.
- Nombre des téléchargements du document (sans distinction du format – PDF, Txt, Doc, HTML).

Il s'agit de statistiques cumulatives qui ne permettent pas d'établir un historique ou l'évolution de l'utilisation. Pour toute autre analyse, il faut aller dans les métadonnées et

faire le tri, comme par exemple l'analyse de HAL-UBO [BERTIGNAC & GAC, 2009] qui sépare thèses et articles.

Une approche comparable est l'étude sur l'utilisation de HAL par la communauté scientifique de l'Ecole Centrale de Lyon [SICOT, 2008] qui présente les statistiques par collection (= dépôts d'un laboratoire de l'Ecole Centrale), en indiquant les nombres maximums, minimums et moyens de téléchargements pour les dépôts d'un laboratoire[4].

### *3.2.2. L'usage par type de document*

Nous avons trouvé quatre études qui donnent quelques informations différenciées sur l'usage des dépôts par type de document (taux de consultation), avec des données et périodes assez hétérogènes et peu comparables.

Dans les archives des écoles d'ingénieurs de Toulouse (OATAO), le taux de consultation des articles se situe à 99%, celui des thèses à 100%. Le nombre moyen de consultations par article est 49, celui par thèse est de 109 [MALOTAUX, 2009].

Pour HAL-UBO, « les thèses sont aujourd'hui les documents (…) les plus consultés » ; elles représentent 17 des 25 documents les plus consultés [BERTIGNAC & GAC, 2009].

Le site de ParisTech[5] contient des informations sur le « TOP 20 des thèses en ligne » dans PASTEL, en termes de téléchargement et fréquence (téléchargement moyen par jour depuis la mise en ligne).

L'analyse de l'utilisation de l'archive institutionnelle de l'IFREMER fournit le nombre mensuel moyen de téléchargements par thèses, rapports et publications (= articles) ; pour les articles, elle prend en compte de l'année de publication (avant vs. après 2000) [MERCEUR, 2007, 2009].

Il est intéressant de noter que ces quelques statistiques confirment sans exception l'intérêt de la littérature grise dans ces archives.

### *3.2.3. L'analyse des fichiers log*

Sur Internet, nous avons trouvé les traces de sept établissements qui analysent les fichiers log de leurs archives ouvertes (TeLearn, OATAO, Institut Jean Nicod, INP Toulouse, CNUM-CNAM, IFREMER, ParisTech). Voici une synthèse.

(1) Outils : Les établissements utilisent beaucoup d'outils différents qui ne fournissent pas toujours le même type d'information. Entre autre, nous avons identifié Google Analytics / Sitemap, Webalizer Xtended, AWStats[6], PhpMyVisite, Analog. Parmi ces outils, on trouve

[4] Ces statistiques par collection (laboratoire) laissent penser qu'il pourrait y avoir une sorte d'effet de masse critique, autour de 100 à 200 dépôts. En dessous de ce seuil, l'utilisation est très limitée. Au-dessus, les chiffres sont plus significatifs.

[5] http://pastel.paristech.org/apropos/stat.html

[6] AWStats, un logiciel open source et gratuit, est également utilisé par le CLEO pour les statistiques

des logiciels libres et commerciaux, utilisés en ligne ou installés en local, à usage professionnel (interne) ou public (en ligne).

Tous ces outils offrent la possibilité d'analyser d'une part, le chemin d'accès d'une consultation (« amont ») et d'autre part, le comportement d'un utilisateur sur site. Quelques exemples.

(2) Accès sur le site : A partir des adresses IP, ces logiciels donnent une information sur la provenance (pays) des visiteurs[7] et de leur configuration (système d'exploitation, navigateur, plugin...), en identifiant le trafic occasionné par les robots. Ceci permet également de différencier les visiteurs uniques, les nouveaux visiteurs, les visiteurs connus, le taux de retour, le nombre de visites par visiteur (cf. [MIN et al., 2008] pour TeLearn ou les statistiques sur le site de ParisTech-PASTEL). En même temps, les traces laissées par les consultations témoignent de la réussite ou de l'échec d'une consultation et du chemin d'accès, si un utilisateur est arrivé directement sur le site, via un site référent (lien) ou un moteur de recherche, et lequel. Les quelques données en ligne sont très claires : la plupart des consultations se fait via Google ; l'accès direct ou à partir de sites référents ou d'autres moteurs de recherche (Yahoo etc.) reste marginal[8]. Pour Archimer, "90% des déchargements sont réalisés à partir des moteurs de recherches standards,Google notamment. Un document indexé par Google sera donc en moyenne déchargé 10 fois plus souvent que les autres." [MERCEUR, 2007]

(3) Comportement sur site : On trouve surtout des informations sur le temps passé sur le site (temps moyen de visite) et sur les documents consultés (accès au texte intégral, analyse des thèmes et domaines). D'autres éléments : comptage des « visites actives », le taux de rebond, le nombre de pages vues, aussi par visiteur, le taux de visites à une page, tout cela cumulatif avec des moyennes journalières ou mensuelles. On trouve aussi des informations de suivi du visiteur (page d'entrée, page de sortie, page de visite d'une seule page).

Il s'agit d'une information riche et détaillée mais très hétérogène, peu exploitée, mal définie.

## *4. Discussion*

Le point de départ de notre étude – la sélection d'archives ouvertes à partir de répertoires et sites de référencement d'autres organismes – induit une certaine hétérogénéité dans la mesure où il n'existe pas de définition a priori d'une archive ouverte. Le résultat est qu'on trouve dans l'échantillon aussi bien de « vraies » archives institutionnelles que de sites à

---

d'utilisation de la plate-forme Revues.org.

[7] Pour OATAO entre avril 2008 et février 2009 : 57% France, avec Etats-Unis et Union Européenne 80% [MALOTAUX, 2009].

[8] Pour OATAO: Google 79%, autres moteurs de recherche 7%, accès direct 6%, sites référents 8% [MALOTAUX, 2009]. Pour le site CNUM-CNAM, le chiffre est de 85% pour Google [BERNARDONI, 2008]. Archimer (juin 2009) : 81% Google, 4% Google Scholar, 4% Archimer, 1% Ifremer Search [MERCEUR, 2009].

caractère patrimonial, de bibliothèques numériques ou de sites web.

Un autre inconvénient de notre approche est le décalage entre la mise en ligne d'une archive ouverte et son référencement. Comme en 2008, nous sommes conscients qu'il existait au moment de la collecte des données d'autres archives ouvertes en France qui ne figuraient pas encore dans les répertoires.

La collecte de données sur le Web pose plusieurs problèmes. Quasiment aucun site ne renseigne sur la date de création de l'archive ou sur son évolution chiffrée. Nous avons été confrontés aux différences de qualité de signalement des métadonnées. Décompter les différents types de documents fut parfois un véritable parcours de combattant. Par exemple, l'interface de recherche d'une archive propose comme principal critère le caractère « publié » ou « non publié » d'un document, une autre classe uniquement par auteur l'ensemble des publications ; pour connaître le nombre exact des différents types de documents, la seule solution possible est de les compter « à la main ».

Pour l'archive nationale HAL, nous avons une image imprécise de la nature des différents documents car un nombre important est classé sous la catégorie « autres ». A l'inverse, l'archive CemOA du CEMAGREF[9] offre une interface de recherche très fouillée : les critères de recherche permettent de filtrer jusqu'à 32 types de documents différents, classés parmi 14 équipes de recherche. La recherche peut afficher les documents publiés au cours d'une période définie et classer les résultats selon l'ordre chronologique croissant ou décroissant.

La récolte des statistiques d'usage s'est faite de façon empirique, en recherchant sur le Web. Nous avons découvert une information très hétérogène, des données brutes d'usage en libre accès sur Internet, des diaporamas présentant des données d'usage, des rapports. Il est probable que d'autres informations appartiennent au domaine de la littérature grise ou se trouvent dans les profondeurs du Web et restent donc pour l'instant inexploitées.

La part importante des notices bibliographiques nous amène à relire la définition d'une archive ouverte. L'analyse du contenu des archives révèle une déviation de l'objectif principal d'une archive. Préoccupés par un souci d'évaluation de l'impact, certains établissements scientifiques donnent la priorité au signalement plutôt qu'à l'accès direct aux documents.

L'importance des archives à caractère patrimonial est une autre limite au concept d'archive ouverte. Au nombre de 30 en 2009, elles représentent 20% des archives et 46% des documents. Or initialement, les archives ouvertes ont été créées pour faciliter et accélérer la communication scientifique au sein des communautés et disciplines. Cette priorité donnée à la valorisation des collections académiques, le mélange entre production scientifique et chargement de vieux fonds numérisés pose un double problème, celui de la cible et de l'objectif du service, et celui de la pertinence des résultats de recherche dans les archives.

[9] http://cemoa.cemagref.fr

## 5. Les statistiques d'utilisation des archives ouvertes : projets de normalisation, recommandations

L'analyse empirique de l'utilisation des archives ouvertes se trouve dans une situation comparable à celle des bibliothèques numériques à leur début : il y a des données, certes, mais elles sont incomplètes, mal définies, partiellement incompatibles et mal diffusées. Et comme pour les revuesdu projet COUNTER, il faudra une prise de conscience par les organismes « clients » et « serveurs » des archives – de l'importance de la maîtrise des données d'usage mais aussi de l'intérêt de leur compatibilité (caractère normatif) et de leur divulgation. Voici à titre d'exemple quelques projets et initiatives, suivis d'une ébauche de recommandations.

### 5.1. PIRUS (JISC)

Le projet anglais "Publisher and Institutional Repository Usage Statistics" (PIRUS) du JISC, dans sa 2e phase depuis 2009, poursuit l'objectif de définir des statistiques d'utilisation des archives ouvertes au niveau d'un article. [BEVAN & NEEDHAM, 2009] Ses principales caractéristiques sont :

Les statistiques doivent être utiles aussi bien pour les auteurs (chercheurs) que pour les institutions, dans une perspective d'évaluation scientifique.

Il s'agit de statistiques pour l'utilisation d'articles, en s'appuyant sur le Digital Object Identifier (DOI) comme identifiant unique. Ceci réduit à priori l'intérêt pour d'autres contenus des archives.

Les statistiques doivent être compatibles avec les rapports du projet COUNTER. PIRUS a ainsi défini un Article Report 1 comme le nombre d'accès réussis au texte intégral d'un article par mois et DOI (*Number of Successful Full-Text Article Requests by Month and DOI*). PIRUS a développé un prototype en format XML pour la collecte et diffusion de ces statistiques. En 2010, PIRUS est censé définir plusieurs « Article Reports », toujours sur le modèle de COUNTER (*core set of standard usage statistics reports*).

Un autre objectif est de développer, après le prototype de la 1e phase, un ou plusieurs logiciel(s) Open Source pour la génération et le partage des statistiques d'utilisation au niveau article (item) dans les archives ouvertes, couplé à une analyse financière des coûts de l'implémentation d'un tel service.

### 5.2. OA Statistik (DINI)

L'équipe du projet allemand OA Statistik[10] développe des outils de transfert et de diffusion de statistiques d'utilisation issues d'archives institutionnelles. Le concept correspond au travail en réseau avec moissonnage des données locales et restitution des statistiques sous forme de services à valeur ajoutée. Les statistiques sont élaborées au niveau du document

[10] http://www.dini.de/projekte/oa-statistik/english/project-results/

déposé (article).

Les statistiques sont destinées aux auteurs pour le suivi d'usage de leur publication, aux lecteurs-chercheurs pour une information sur la pertinence du document et pour la création d'alertes, et aux institutions comme contribution à l'évaluation de l'impact de leur production scientifique.

Le concept et l'architecture de ce projet nécessitent une forte normalisation afin d'assurer l'interopérabilité entre toutes les composantes du réseau. A titre d'exemple, OA Statistik produit non pas un mais trois indicateurs de téléchargement, à partir de trois différentes définitions et méthodes de comptage (COUNTER, IFABC, LogEc).

### *5.3. IFREMER*

Comment le format des dépôts peut-il impacter le référencement par les moteurs de recherche (surtout Google) ? Une analyse de l'IFREMER formule des conseils pour optimiser la visibilité des dépôts et, indirectement, leur utilisation : déposer un seul fichier avec toutes les (méta)données, renseigner les propriétés du format PDF, réduire la taille des fichiers, etc. [MERCEUR, 2009]. Même si l'analyse peut paraître un peu en marge des autres projets, elle y appartient dans la mesure où elle fait le lien entre données d'utilisation et document, et qu'elle contribue à une certaine normalisation des dépôts.

### *5.4. D'autres initiatives et projets*

Dans le domaine des statistiques d'utilisation, il y a une synergie évidente entre les projets comme COUNTER, PIRUS et OA-Statistik. L'impact du JISC anglais avec des projets structurants comme Usage Statistics Review [MERK et al., 2008] et IRS[11] est certain. Néanmoins, d'autres projets contribuent au processus de normalisation. Voici quatre exemples :

**Publishing and the Ecology of European Research (PEER) :** Il s'agit d'une étude sur l'impact des archives ouvertes sur le modèle économique traditionnel de l'édition scientifique. Un des premiers travaux fut la définition d'un format commun des fichiers log (common logfile format) afin de pouvoir intégrer des données en provenance de différentes sources – une approche comparable à celle du projet DINI [BAYER-SCHUR et al., 2009].

**RePEc :** L'archive ouverte en sciences économiques RePEc produit des données d'utilisation d'une grande qualité.[12] Leur service LogEc exploite plus de 45 millions téléchargements depuis 1998, une expérience incomparable avec un effet certain sur le développement d'autres systèmes, services et définitions.

**IFABC :** Parmi les indicateurs recommandés par l'organisation "International Federation of Audit Bureaux of Circulations"[13] figure aussi une définition d'indicateurs d'usage d'un

[11] http://irs.eprints.org/

[12] http://logec.repec.org/

[13] http://www.ifabc.org/

site web, avec une terminologie normalisée pour « utilisateur », « visite » etc.

**SURF :** La fondation néerlandaise SURF finance un projet de normalisation, « Statistics on the Usage of Repositories » (SURE)[14], entre autre pour faciliter l'agrégation des fichiers log issus des archives ouvertes.

### *5.5. Recommandations*

Pour progresser dans l'analyse des statistiques d'utilisation des archives ouvertes, nous avons dressé une liste de six recommandations qui s'appuie sur notre enquête et sur l'analyse des autres projets et initiatives.

1. Destinataires : La diffusion des statistiques d'utilisation doit inclure les auteurs des dépôts, les utilisateurs (visiteurs) du site et l'institution responsable du contenu du site. Ceci ne veut pas dire que tout le monde a besoin des mêmes informations ; mais vu le caractère spécifique et l'environnement des archives ouvertes, il est important de créer un contexte de transparence qui correspond à la philosophie de l'accès libre et du Web 2.0.
2. Principe COUNTER : Les statistiques d'utilisation des archives doivent suivre l'approche du projet COUNTER, c'est-à-dire, il faut définir plusieurs niveaux de statistiques et d'indicateurs, avec un niveau de base assez simple, équivalent au Journal Report 1[15] (« Archive Report 1 »).
3. Fichiers log : De même, il faudra déterminer un nombre restreint d'éléments minimum pour exploiter les fichiers log. Ces éléments devraient permettre d'identifier le visiteur (qui), décrire l'objet (quoi) et le type de sa requête, préciser la date et la durée de la visite (quand), et attribuer un identifiant unique à chaque visite.
4. Dictionnaire : Il faut recenser les concepts, termes et données clés de l'analyse (consultation, visite, téléchargement, accès, *request*, *hit*…), dresser un glossaire et définir ces termes sur le modèle du projet COUNTER. Ce sera nécessairement en deux langues (anglais/français), vu que la plupart d'outils et initiatives ont recours à l'anglais. Il y aura donc aussi un travail de traduction[16].
5. Périodicité : Les statistiques devraient être établies sur la base d'une période mensuelle, avec un cumul annuel, et disséminées dans les 30 jours après la fin du mois.
6. Texte intégral : Les statistiques d'utilisation doivent différencier l'accès à un

[14] http://www.surffoundation.nl/nl/projecten/Pages/SURE.aspx

[15] Le Journal Report 1 (JR1) chiffre le nombre de requêtes réussies d'un article en texte intégral, par revue et par mois. Chaque revue est identifiée par son titre et ISSN (print et/ou online). Le tableau contient également le cumul annuel pour chaque revue et le cumul mensuel pour l'ensemble des titres. La 2e version du JR1 ajoute le nom d'éditeur et la distinction HTML/PDF.

[16] Cf. le glossaire anglo-français pour COUNTER http://counter.inist.fr/sites/counter/IMG/pdf/COUNTER_-_Code_de_Bonnes_Pratiques_v2a.pdf

document en texte intégral et celui à une notice, surtout quand elle n'est pas accompagnée d'un document. Il y aura donc au moins trois niveaux : accès aux métadonnées d'un document, accès aux documents, accès à une notice sans document.

Dans un 2e temps, nous recommandons de divulguer les statistiques d'utilisation des archives ouvertes dans un environnement de services à valeur ajoutée, sur le modèle de la PloS[17] ou du projet DINI. Cet environnement pourrait inclure :

1. Des statistiques modulables par l'utilisateur, soit en ligne, soit après téléchargement des tableaux (statistiques par collection, type de documents, période etc.).
2. Des tableaux synthétiques (statistiques annuelles, tableaux comparatifs, cumul par année de publication, type de document, collection, domaine) ; dont le "average lifetime usage per journal".
3. Une assistance technique et une explication de toutes les données et statistiques, par exemple sous forme d'une liste de questions-réponses (FAQ).
4. Ajout d'autres outils pour mesurer l'impact d'un document déposé comme les citations, les liens, les annotations, le social tagging et le bookmarking etc… [BATH, 2009]

Il ne s'agit pas nécessairement d'inventer de nouveaux outils mais plutôt d'adapter les outils d'autres projets existants à l'environnement français [CARR et al., 2008].

## *6. Conclusion*

Notre communication décrit la méthode et les premiers résultats de notre étude sur le développement et l'usage des archives ouvertes en France. Une autre communication ajoute quelques éléments sur la littérature grise [SCHÖPFEL et al., 2009]. Voici la suite de nos travaux :

Dans un premier temps, nous finirons les analyses statistiques de la base de données des archives ouvertes, en mettant l'accent sur le développement, le contenu et quelques aspects qualitatifs.

En même temps, nous mènerons l'étude de cas du site IRIS de l'université Lille 1 (USTL) à partir des fichiers log.

Ces études aboutiront à la production d'un rapport final qui contiendra aussi nos recommandations pour les études des statistiques d'utilisation des archives ouvertes, avec un début de travail sur la terminologie. Ce rapport sera mis en ligne et librement accessible.

En marge de notre projet, nous travaillerons avec une BU allemande sur le référencement des archives ouvertes par les moteurs de recherche pour mieux comprendre le lien entre référencement et utilisation.

Nous allons également renforcer nos liens avec le JISC/PIRUS et le DINI pour mieux nous positionner dans le paysage international. D'autres propositions de coopération sont à

[17] Cf. PLoS http://article-level-metrics.plos.org/

l'étude (Russie, Japon).

Au niveau national, nous allons poursuivre notre collaboration avec COUPERIN dans le domaine de l'analyse des usages des ressources en ligne, en essayant de mieux comprendre le comportement des communautés scientifiques et d'accompagner les bibliothèques et institutions scientifiques dans le choix de leur politique d'archive ouverte.

## *7. Bibliographie*


ARMBRUSTER, C., ROMARY, L., (2009). *Comparing Repository Types: Challenges and Barriers for Subject-Based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication* (November 23, 2009). http://ssrn.com/abstract=1506905

AUBRY, C., JANIK, J., (2005). *Les Archives Ouvertes, enjeux et pratiques : guide à l'usage des professionnels de l'information*. Paris : ADBS

BATH, M. H., (2009). Open access repositories in computer science and information technology: an evaluation. *IFLA Journal*, vol. 35, n° 3, p. 243-257.

BAYER-SCHUR, B., et al. (2009). *Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers*. Technical report, PEER Project..

BERTIGNAC, C., GAC, D., (2009). La voie verte : les archives ouvertes. In *Open Access Week. Institut Universitaire Européen de la Mer, Brest, 19 octobre 2009*.

BEVAN, S., NEEDHAM, P., (2009). Repository usage statistics - can you count on them? In *ETD 2009. 12th International Symposium on Electronic Dissertations and Theses. University of Pittsburgh, June 10-13, 2009*.

BOUKACEM, C., SCHÖPFEL, J., (2005). Statistiques d'utilisation des ressources électroniques en ligne : le projet COUNTER. *Bulletin des Bibliothèques de France*, vol. 50, n° 4, p. 62-66.

BOUKACEM-ZEGHMOURI, C., SCHÖPFEL, J., (2008). On the usage of e-journals in French universities. *Serials*, vol. 21, n° 2, p. 121-126.

CARR, L., BRODY, T., SWAN, A., (2008). Repository statistics: What do we want to know? In *Third International Conference on Open Repositories, 1-4 April 2008*.

CREPPY, R., (2007). Archives ouvertes, archives institutionnelles et protocole français. *Bulletin des Bibliothèques de France*, vol. 52, n° 6, p. 42-45.

GOUAT, I., (2009). Dépôt et comptage des publications du LIRMM extraites de HAL dans le cadre des évaluations des chercheurs/labo. In *Journée d'étude : Indicateurs bibliométriques, production scientifique et évaluation des chercheurs, Grenoble. 3 avril 2009.*

MALOTAUX, S., (2009). OATAO Archive ouverte multi-établissements. Bilan après un an d'existence. In *Journées d'études sur les archives ouvertes. Consortium COUPERIN. Paris. 2 et 3 avril 2009.*

MERCEUR, F., (2007). Gestion d'une archive et d'un moissonneur, l'exemple de

l'IFREMER. In *RPIST 2007, 20 juin 2007, Nancy*.

MERCEUR F., (2009). Optimiser la visibilité de vos dépôts, au vu des statistiques d'Archimer. In *Open Access Week. Institut Universitaire Européen de la Mer, Brest, 19 octobre 2009.*

MERK, C., SCHOLZE, F., WINDISCH, N., (2009). Item-level usage statistics: A review of current practices and recommendations for normalization and exchange. *Library Hi Tech,* vol. 27, n° 1, p. 151-162.

MIN, S., BALACHEFF, N., ZEILIGER, J. (2008) TeLearn, une archive ouverte multilingue dans le domaine des technologies pour l'apprentissage. *AMETIST*, n° 2, partie 3.

SCHÖPFEL, J., (2008). Le projet COUNTER. Eléments-clés et actualités. In *Journée d'étude : Ressources électroniques dans les bibliothèques : Mesures et Usages. Université Charles de Gaulle Lille 3, 28 novembre 2008.*

SCHÖPFEL, J., PROST, H., (2009). L'accès libre en mouvement. Journées d'étude sur les archives ouvertes du consortium COUPERIN, 2-3 avril 2009. Texte non publié.

SCHÖPFEL, J., BOUKACEM-ZEGHMOURI, C., PROST, H., (2009). Usage of grey literature in open archives. In: *Eleventh International Conference on Grey Literature GL11. The Grey Mosaic: Piecing It All Together. Washington D.C., December 14-15, 2009.*

SCHÖPFEL, J., STOCK, C., (2009a). Grey literature in French Digital Repositories: A Survey. *The Grey Journal*, vol. 5, n° 3, p. 147-161.

SCHÖPFEL, J., STOCK, C., (2009b). Les archives ouvertes en France – Un potentiel documentaire pour la formation à distance. *Distances et Savoirs*, vol. 7, n° 4, p. 443-456.

SICOT, J., (2008). Bilan de l'utilisation de HAL par la communauté scientifique de l'Ecole Centrale de Lyon. In *Conseil Scientifique du 10 juillet 2008.*

TARABORELLI, D., (2005). *Institutnicod.org. Rapport sur l'impact du site sur la visibilité du laboratoire (2001-2005).* 17 mai 2005.

VAN DER GRAAF M., VAN EIJNDHOVEN, K., (2008). *The European Repository Landscape. Inventory Study into the Present Type and Level of OAI-Compliant Digital Repository Activities in the EU*. Amsterdam University Press.

WILLINSKY, J., (2006). *The access principle: the case for open access to research and scholarship*. Cambridge, Mass., MIT Press.