Développement et Usage des Archives Ouvertes en France. 2e partie : Usage

Joachim Schöpfel, Hélène Prost

Université Charles-de-Gaulle Lille 3
Laboratoire GERIICO
Groupe d'Etudes et de Recherche Interdisciplinaire en Information et Communication

Développement et Usage des Archives Ouvertes en France

Rapport

2e partie : Usage

Joachim Schöpfel, Université Lille 3
Hélène Prost, INIST-CNRS

Lille, octobre 2010

Résumé : Le rapport présente les résultats d'un projet de recherche mené en 2009 à l'université Charles-de-Gaulle Lille 3. L'objectif du projet : évaluer les résultats de la politique en faveur des archives ouvertes en France. La 2e partie du rapport intitulé « Usage » contient un état de l'art de l'analyse des statistiques d'utilisation des archives ouvertes et fournit quelques éléments chiffrés sur les archives ouvertes en France, à partir de données collectées en ligne sur plusieurs sites. L'enquête est suivie d'une étude de cas, l'analyse des fichiers log de l'archive institutionnelle IRIS de l'université Lille 1.

Mots clés : archives ouvertes, information scientifique, publication scientifique, accès libre, statistiques d'utilisation.

Abstract : The report contains the results of a research project conducted in 2009 at the university Charles-de-Gaulle Lille 3. The objective of the project: evaluate the results of the policy of open repositories in France. The 2nd part of the report contains a state of the art on the analysis of usage statistics of open archives and produces some figures on French open archives, based on data collected on line, on different web sites. The survey is followed by a case study, a log file analysis of the institutional repository IRIS from the university of Lille 1.

Keywords : open repositories, scientific information, academic publishing, open access, usage statistics.

Contact : Joachim Schöpfel joachim.schopfel@univ-lille3.fr

Financement : Projet BQR 2009 Université de Lille 3

Remerciements : Le projet a été subventionné en 2009 par le service recherche de l'université de Lille 3 que nous remercions pour son conseil et son aide logistique. Nous remercions également la direction et l'équipe du SCD de Lille 1 pour leur soutien ainsi que tous les participants – enseignants-chercheurs, professionnels et étudiants - ayant contribué au succès de cette étude. Nos remerciements notamment à Isabelle Le Bescond pour une relecture attentive du manuscrit.

Information sur le projet

Acronyme	DUAO-F
Nom	Développement et usage des archives ouvertes en France – une étude empirique
Date début	1er mars 2009 **Date fin** 31 décembre 2009
Etablissement porteur	Université Charles de Gaulle Lille 3
Coordinateur projet	Joachim Schöpfel
Coordonnées	Laboratoire GERiiCO, Domaine Universitaire du Point de Bois, BP 60149, 59653 Villeneuve d'Ascq Cedex Email: joachim.schopfel@univ-lille3.fr Tél.: 0320 416 153 / 0688 350 147
Partenaires	Université Charles de Gaulle Lille 3 Laboratoire GERiiCO équipe Savoirs, Information, Document (SID) Chérifa Boukacem-Zeghmouri (MCF) Email boukacemc@yahoo.fr Tél.: 0620 621 812 Aude Sauer-Avargues (étudiante) Email saueravargues.aude@free.fr Mickaël Malandran (étudiant) Email mickael.malandran@laposte.net Université des Sciences et Technologies de Lille Service Commun de Documentation Julien Roche (directeur) Email julien.roche@univ-lille1.fr Tél. : 0320 434 410 / 0320 337 199 Isabelle Le Bescond Email Isabelle.Le-Bescond@univ-lille1.fr François Lefebvre Email francois.lefebvre@univ-lille1.fr Université d'Amsterdam DAREnet Saskia Woutersen-Windhouwer (Specialist Electronic Publishing & Repository Manager) Email S.Windhouwer@uva.nl Tél.: 0031 20 525 Institut de l'Information Scientifique et Technique du CNRS Hélène Prost (chargée de ressources documentaires) Email prost@inist.fr Tél.: 0383 504 600 ou 0672 427 630
Site web	http://usageao2009.pbworks.com/ Liste de diffusion duao@univ-lille1.fr
Programme de recherche	Bonus Qualité Recherche 2009 Université Lille 3
Coordination programme	Michel Crubellier

Table des matières

Liste des figures et tableaux

Synthèse

La France figure parmi les pays fortement engagés dans le mouvement vers le libre accès à l'information scientifique, par le biais de la communication scientifique directe, c'est-à-dire la mise en place d'archives ouvertes sur Internet et la création de revues gratuites en ligne. Néanmoins, à ce jour et contrairement à d'autres pays, il n'existe que peu d'études empiriques sur les résultats de cet investissement public. Notre étude tente de contribuer à l'évaluation du développement des archives ouvertes en France. La première partie du rapport final a fourni des éléments sur la typologie, la taille, le contenu et le développement des archives ouvertes. Leur utilisation par les communautés scientifiques fait l'objet de cette 2e partie du rapport.

L'étude inclut un état des lieux de l'analyse des statistiques d'usage des archives ouvertes, et plus particulièrement de la méthode de l'analyse des fichiers log. La partie empirique poursuit une double approche : une enquête menée en 2009 sur un échantillon quasi-exhaustif des sites français, à partir d'information et de données publiées et/ou disponibles en ligne ; et une étude de cas du site IRIS de l'université de Lille 1.

L'état des lieux présente quelques projets majeurs dont PIRUS et OA-Statistik et décrit le cadre méthodologique des équipes CIBER et Cybermetrics dans le domaine des fichiers log.

Les résultats de l'enquête sont les suivants : Nous avons trouvé des statistiques pour dix sites. C'est peu représentatif, comparé au nombre total (= 7%). Cette information porte sur l'accès ou le téléchargement des documents par auteur ou institution, sur le type de documents, et/ou sur l'analyse détaillée des fichiers log, en séparant l'accès et le comportement sur le site.

L'étude de cas IRIS est menée selon cette procédure : L'analyse inclut tous les fichiers log de 2009. A partir du cadre conceptuel de CIBER, les données statistiques ont été générées avec le logiciel Urchin. L'étude porte sur neuf indicateurs d'activité (dont le nombre des consultations ou visites, téléchargements et sessions, la profondeur de visite et la durée d'une session), huit indicateurs de la recherche d'information (dont le titre et l'année de publication des documents consultés, la taille des fichiers téléchargés, le mode de navigation et le chemin d'accès) et trois données concernant les utilisateurs (dont l'origine géographique). Le rapport présente chaque indicateur avec les résultats de l'analyse des fichiers log 2009 d'IRIS.

La discussion porte sur cinq points : sur la dissémination des statistiques d'utilisation en France, sur l'intérêt et les limites de l'analyse des fichiers log, sur les indicateurs d'activité ou trafic (consultations et téléchargements), sur l'analyse du comportement sur site (recherche et navigation), et sur l'information concernant les utilisateurs.

La dernière partie du rapport formule une série de recommandations pour développer l'analyse des statistiques d'utilisation des archives ouvertes, pour développer des services à valeur ajoutée, et pour développer l'analyse des fichiers log.

Le projet vient renforcer et compléter les projets en cours de l'équipe Savoirs, Information, Document (SID) du laboratoire GERIICO (université Lille 3), en termes d'objets d'études, de méthodologie et de résultats. Il soutiendra le développement du positionnement et des partenariats du laboratoire GERIICO au niveau national et européen. A ce jour, il a donné lieu à plusieurs communications et publications (cf. annexe A).

1. Introduction

Dans l'environnement du mouvement vers le libre accès 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. Aubry & Janik, 2005; Willinsky, 2006; Lutz, 2009, Björk et al., 2010). 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 ne sont pas seulement utiles à la recherche mais sont devenus 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.

Mesurer l'usage des archives ouvertes est un moyen d'évaluer leur impact dans le paysage de l'information scientifique. Les statistiques d'usage intéressent tous les acteurs concernés pour les raisons suivantes (cf. Björnshauge, 2006 et Carr et al., 2008) :

1. Les chercheurs-auteurs sont intéressés par le suivi en ligne de l'utilisation de leurs publications et par les centres d'intérêts des lecteurs.[1]

2. Les hébergeurs veulent connaître l'utilisation des documents pour analyser la valeur de la mise en ligne de ces contenus et pour évaluer le rapport coût-bénéfice de l'investissement.

3. Les organismes de recherche souhaitent obtenir de l'information sur l'intérêt des dépenses dans les archives ouvertes et sur l'impact de "leurs" publications.

4. Les agences de moyens (= organismes de financement) cherchent d'autres méthodes quantitatives et transparentes pour évaluer la performance et l'impact des projets scientifiques qu'elles subventionnent. Elles demandent des comptes, des statistiques d'utilisation, des indicateurs de performance et d'impact, de la qualité.

Les professionnels exigent une information précise sur l'utilisation des ressources en ligne qu'ils achètent aux éditeurs. Ils veulent pouvoir déterminer les tendances des usages, comprendre l'origine et la nature de l'utilisation des dépôts et développer des indicateurs.[2] L'analyse de leurs coûts et la recherche d'un modèle économique durable pour les archives ouvertes sont à l'ordre du jour.

De même, les statistiques d'utilisation peuvent servir de base pour développer de nouveaux services à valeur ajoutée pour les auteurs (évaluation de leur impact via l'utilisation de leurs publications, ponctuellement ou en continu) et surtout pour le chercheur-utilisateur. Deux exemples : la mise en place d'alertes (= les articles, travaux etc. les plus consultés, cf. LogEc/RePEc[3]), et le calcul de la pertinence des réponses dans une interface de recherche (Metje, 2009).

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 peu diffusées. Une étude récente sur l'édition SHS en France (GFII, 2009) a regretté l'absence de visibilité sur l'impact

[1] Le lien entre l'usage des articles mis en ligne dans une archive ouverte et leur citation (impact) a été établie depuis plusieurs années (Brody, 2006 ; Schimmer, 2006 ; Davis & Fromerth, 2007).
[2] Cf. le projet PEER (Fry et al., 2009).
[3] http://logec.repec.org/

de l'activité des archives ouvertes. L'analyse empirique de l'activité en ligne est encore l'exception. Dans l'enquête du consortium COUPERIN, seulement six projets sur 74 (= 8%) avaient produit des statistiques de consultation (Bruley et al., 2007). Pourtant, les rares résultats semblent plutôt prometteurs : "La fréquentation indiquée se situe entre 200 et 130 000 consultations moyennes. Même si le nombre de documents présents dans les archives n'est pas forcément très important, le nombre de consultations pour les projets qui ont fourni l'information lui est élevé" (*loc. cit.*).

Comme pour l'offre des revues, *ebooks*, bases de données etc. de la part des éditeurs scientifiques il y a quelques années, 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.

1.1. Statistiques d'utilisation : projets et réalisations

Dans le domaine des statistiques d'utilisation, il y a une synergie évidente entre plusieurs projets. L'impact du JISC anglais avec des projets structurants comme Usage Statistics Review (Merk & Windisch, 2008) et Interoperable Repository Statistics (Carr & Brody, 2006) est certain[4]. Le projet IRS s'intéresse aux statistiques de téléchargement des archives institutionnelles à partir des fichiers log[5] et pose la question de l'agrégation et de la diffusion de ces données ; l'équipe projet s'est entourée d'un comité consultatif composé par des représentants des grandes archives internationales (ArXiv, CERN, CNRS, RePEc, DARE...).

Le projet IRS exploite des fichiers log des systèmes Eprints et DSpace et produit des statistiques par dépôt, catégorie, auteur ou thématique etc., avec une visualisation des résultats sous forme de graphiques et un filtrage du trafic des robots[6]. La JISC Usage Statistics Review propose un cadre pour l'agrégation des fichiers log des archives ouvertes afin de produire des statistiques d'utilisation comparables, en stipulant un certain nombre de données minimales (cf. 1.2.).

Le projet COUNTER[7] témoigne d'un intérêt grandissant pour les statistiques d'usage d'articles individuels (Shepherd, 2009). Le problème est que l'article peut se trouver sur plusieurs sites en même temps. L'enquête de PIRUS auprès des éditeurs et l'analyse de plusieurs systèmes d'archives institutionnelles[8] affichent un DOI plus ou moins accepté comme identifiant unique pour les articles. Néanmoins, il n'existe pas de procédure standard pour attribuer des DOI et il n'y a pas que d'articles dans les archives mais aussi d'autres types de documents.

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

[4] http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/usagestatisticsreview.aspx
http://irs.eprints.org/

[5] Notre rapport utilisera systématiquement le terme de « fichier log » à la place de « fichier journaux ».

[6] Citation de Carr et al. (2008): "The range of analyses encompasses simple access counts (how many times was this item downloaded), league table functionality (what are the top 10 most downloaded items / authors) and more sophisticated calculations (which items are the 'highest climbers' in the league tables). A simple deployment of the statistics might be to include a monthly download graph on each item's abstract page; a more extensive use of the package could make an aggregate collection of statistics available from different perspectives. One such pre-prepared example is provided by IRStats – the so-called "Download Dashboard" (…) intended to give the author of an item a comprehensive picture of its accesses and the reason for those accesses."

[7] Cf. http://counter.inist.fr/ et Boukacem-Zeghmouri & Schöpfel (2005).

[8] DSpace, Eprints, Fedora, Digital Commons

des archives ouvertes au niveau d'un article (Bevan & Needham, 2009)[9]. Le projet s'adresse aux auteurs (chercheurs) et aux institutions, dans une perspective d'évaluation de la production scientifique.

PIRUS applique les recommandations COUNTER aux dépôts dans les archives institutionnelles, avec une méthodologie mixte (état de l'art, enquête, développement ou "testing"). 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*). D'autres indicateurs sont à l'étude ; le projet prévoit pour 2010 plusieurs autres « Article Reports » de complexité progressive, toujours sur le modèle de COUNTER (*core set of standard usage statistics reports*).

Un autre objectif est de développer, après un premier prototype en XML, un ou plusieurs logiciel(s) Open Source pour la génération, la collecte et le partage des statistiques, avec une interface de type tableau de bord ("dashboard style"). Le développement informatique sera accompagné d'une analyse financière des coûts de l'implémentation d'un tel service.

La limite du projet PIRUS est qu'il étudie uniquement les statistiques d'utilisation d'articles, en s'appuyant sur le Digital Object Identifier (DOI) comme identifiant unique. Ceci réduit à priori l'intérêt pour d'autres types de publication (preprints, thèses, rapports etc.).

La suite du projet sera l'agrégation des statistiques de différents sites ("different copies of the same document"), et l'ajout d'un accès aux métadonnées ("item views") au comptage des téléchargements du texte intégral ("fulltext downloads" cf. COUNTER), afin de faire le lien entre les statistiques et la description bibliographique des articles.

Le réseau allemand DINI[10] a subventionné une étude proche du projet anglais PIRUS. L'équipe OA Statistik[11] 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 déposé (article).

Les statistiques sont destinées non seulement 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, mais aussi 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). D'après les données sur la page de démonstration[12], ces trois statistiques normalisées sont fortement corrélées, avec un coefficient autour de r=.9 ; elles mesurent donc plus ou moins le même phénomène.

Le projet a également élaboré des spécifications pour le format et l'échange des données statistiques, adaptées à plusieurs applications (dont OPUS et DSpace).

L'annexe B présente d'autres projets et réalisations qui, malgré toutes leurs différences, ont plusieurs points communs : ils proposent un cadre terminologique pour les concepts-clés et contribuent à la normalisation des formats et procédures d'extraction et de traitement des

[9] http://www.jisc.ac.uk/publications/reports/2009/pirusfinalreport.aspx
[10] Deutsche Initiative für Netzwerkinformation http://www.dini.de/
[11] http://www.dini.de/projekte/oa-statistik/english/project-results/
[12] http://oa-statistik.sub.uni-goettingen.de/statsdemo/

fichiers log. Certains projets mettent à la disposition des auteurs, des internautes et des institutions des outils spécifiques tels que tableaux de bord, interfaces de consultation et de recherche etc.

1.2. L'analyse des fichiers log

Mesurer l'empreinte d'une institution sur le web recouvre trois approches méthodologiques, l'analyse de son activité sur Internet, l'évaluation de son impact et l'étude de son utilisation, c'est-à-dire du trafic sur son site (Aguillo, 2009). L'analyse des fichiers log contribue à ce dernier objectif.

Comment extraire les données d'un site Web ? Il existe un certain nombre de logiciels sur le marché (Google Analytics, Webalizer Xtended, AWStats etc.) qui produisent des statistiques d'utilisation plus ou moins détaillées. Mais il n'existe pas de norme ou de terminologie acceptée, et la traduction des termes en français paraît parfois assez aléatoire.

Quelle information peut-on produire à partir d'une analyse des fichiers log ? La JISC Usage Statistics Review (Merk & Windisch, 2008) stipule *ad minima* cinq données : l'utilisateur, le contenu (item), l'utilisation (type de consultation), la date (avec l'heure), l'identifiant unique. Aguillo et al. (2010) mentionnent "visits, visitors and downloads" comme contenu des rapports d'usage des archives ouvertes.

La réduction à cinq données paraît exagérée, au regard du potentiel des fichiers log. Le schéma cadre utilisé par l'équipe Cybermetrics propose 16 critères opérationnels regroupés en trois catégories générales, popularité, comportement et téléchargements (cf. figure 1).

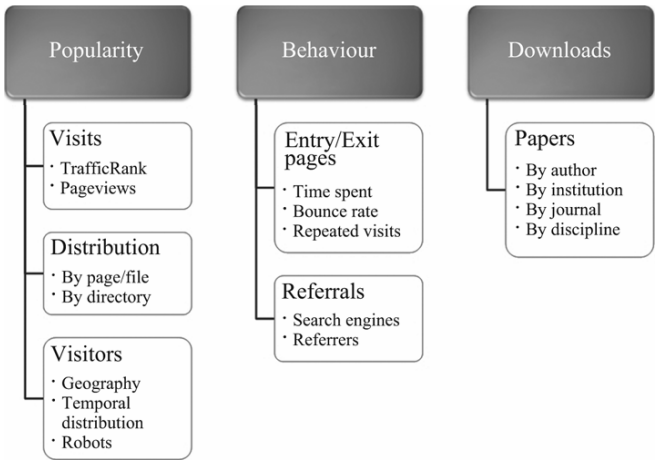

Figure 1 : Schéma cadre de l'analyse des fichiers log (Aguillo, 2009)

Ce schéma contient entre autre le nombre de pages consultées, la répartition du trafic par page ou fichier, l'origine des visiteurs, le chemin d'accès au site et une description détaillée des documents téléchargés. Mais Cybermetrics met en garde : toutes ces données n'ont pas le même intérêt pour une analyse de l'utilisation. En fonction des besoins et objectifs, il faut faire des choix. Et parfois il faut paramétrer le serveur pour obtenir une information utile.

L'équipe anglaise CIBER (Centre for Information Behaviour and the Evaluation of Research) publie depuis plus de cinq ans des études basées sur l'analyse des fichiers log (*deep log analysis*) des ressources documentaires en ligne, des collections et bibliothèques

numériques[13]. CIBER a présenté fin 2009 un schéma global avec une trentaine de critères et d'indicateurs regroupés en trois catégories, « activity metrics », « information seeking characteristics » et « user characteristics » présentés dans le tableau suivant :

Activity Metrics	Information Seeking Characteristics		User Characteristics
	A. Type of content viewed	B. Searching style	
1. Number of pages viewed	1. Number of sources used in a session	1.Search approach adopted	1. Subject/ discipline
2. Number of full-text downloads	2. Names of sources used/not used	2. Number of searches conducted in a session	2. Job status
3. Number of sessions conducted	3. Subject of source	3. Number of search terms used in search	3. Geographical location
4. Site penetration	5. Age of source used	4. Form of navigation	4. Name of organisation
5. Time spent viewing a page	6. Type of material viewed	5. From where users arrive from	5. Type of organization used to access the service
6. Time spent on a session	7. Type of full-text view		6. User demographics:gender, age etc. (if available)
7. Number of searches undertaken in session	8. Size of source used		
8. Number of repeat visits made	9. Publication status of an article		
9. Number of sources used			
10. Number of views per source			

Tableau 1 : Indicateurs pour l'analyse des fichiers log (Nicholas et al., 2009b)

Ce schéma est le fruit de l'analyse de plus de 1,5 millions sessions dont les résultats ont été publiés depuis 2005 (cf. tableau 2). Il procure l'impression d'une synthèse cohérente, d'un cadre logique et raisonné pour la méthode d'analyse des fichiers log.

Publication	Corpus	Mois	Sessions	Pages
Nicholas et al 2005	Blackwell	2	1274960	4 813 806
Nicholas et al 2007	OHIOLink	15	339 041	2 250 196
Nicholas et al 2008	Elsevier	18	16 865	110 029
Nicholas et al 2009a	OUP	12	233 368	631 141

Tableau 2 : Etudes de l'équipe CIBER (2005-2009)

Néanmoins, une comparaison analytique des publications de CIBER ajoute un bémol à cette apparence. Les données de CIBER sont spécifiques et sans lien avec COUNTER ou d'autres études et projets. On ne peut guère parler d'une collecte systématique de statistiques qui s'appuierait sur un concept, sur une théorie ou sur une hypothèse. En fait, CIBER se contente d'exploiter les données mises à disposition par les éditeurs (Blackwell, Oxford University Press, Elsevier) ou établissements (consortium OHIOLink). Or, en absence d'une normalisation, chaque éditeur propose ses fichiers et données spécifiques.

Le schéma de CIBER a deux autres limites ou contraintes : Il ne concerne qu'un seul type de publication, les articles de revues scientifiques (en ceci, CIBER poursuit la même stratégie que PIRUS). Et il intègre certains éléments qui ne proviennent pas des fichiers log mais d'autres sources, notamment des fichiers d'utilisateurs des éditeurs et de leurs listes de titres.

[13] En fait, la première étude de CIBER date déjà de 1999 ; mais elle portait sur un site de presse (*The Times/Sunday Times*) et elle était assez succincte (Nicholas et al., 1999).

Tout cela donne l'impression d'une approche assez exploratoire et empirique sans conceptualisation ou cadre théorique, avec des données et notions variables sans liens entre eux.

Si on met le schéma de CIBER en correspondance avec celui de Cybermetrics, cette impression se renforce (cf. tableau 3). Il est difficile de trouver une cohérence entre les deux approches. Seulement huit données se retrouvent dans les deux schémas. Leur répartition sur les différentes catégories ne suit aucune logique apparente.

	Activity Metrics	Information Seeking Characteristics		User Characteristics
		A. Type of content viewed	B. Searching style	
Popularity	1. Number of pages viewed	2. Names of sources used/not used		3. Geographical location
Behaviour	5. Time spent viewing a page		5. From where users arrive from	
	8. Number of repeat visits made			
Downloads	2. Number of full-text downloads	3. Subject of source		

Tableau 3 : Correspondance entre les deux schémas CIBER et Cybermetrics

Cette absence de convergence sur la typologie et sur la pertinence des données à mesurer laisse à penser que le terrain empirique des fichiers log reste encore assez peu exploré et qu'il ne faut pas attendre une forte conceptualisation des questions et résultats.

1.3. Objectifs de l'étude

Notre propre étude poursuit trois objectifs :

(1) Compléter l'enquête sur le développement des archives ouvertes en France par un volet « usage », à partir de deux hypothèses (cf. Schöpfel & Prost, 2010) :

Il y a davantage d'information sur l'utilisation de ces archives, en termes d'accès (visites) et de téléchargements.

En 2008, il n'y avait pratiquement pas d'information sur l'utilisation réelle des archives ouvertes. Ce n'est pas une particularité française[14]. Notre deuxième hypothèse part donc du principe que généralement, la situation n'a pas bougé depuis 2008 mais compte tenu du nombre de projets et d'initiatives nouvelles, il y a une certaine probabilité de trouver davantage de statistiques d'utilisation qu'avant.

L'information sur l'utilisation n'est ni exhaustive, ni normalisée.

Même constat. Nous n'avons pas connaissance d'initiatives dans le paysage français en faveur d'une généralisation ou d'une normalisation de ces statistiques. Nous partons donc dans l'expectative de trouver des données hétéroclites, non comparables, non représentatives. D'après l'expérience dans d'autres pays, certains chiffres sont diffusés sur le web, d'autres par liste de diffusion, dans un rapport, plus rarement dans un article ou une communication.

(2) Appliquer la méthode de l'analyse des fichiers log à une archive ouverte (IRIS), sous forme d'une étude de cas.

(3) Formuler quelques recommandations concernant les statistiques d'utilisation en général et l'analyse des fichiers log, en particulier.

[14] Bath (2009) a mené une enquête sur des archives ouvertes dans un domaine particulier (informatique, mathématiques) ; une seule archive sur neuf fournit des statistiques d'accès aux utilisateurs (Caltech Computer Science Technical Reports http://caltechcstr.library.caltech.edu/) ; quatre le font sur demande, quatre autres ne proposent rien. Il faut s'attendre à trouver la même situation en France.

La situation actuelle ouvre la voie à une alternative, limitée à quelques données robustes, interprétables, éprouvées et conceptualisées dans un cadre théorique et méthodologique solide.

L'intérêt d'une telle approche est de contribuer à la production et à l'exploitation des statistiques d'utilisation des archives ouvertes en France ; ainsi cette étude favorisera l'évaluation et la compréhension du mouvement du libre accès à l'information scientifique. -

Ajoutons deux remarques au préalable : Notre enquête n'a pas essayé de définir le terme « archive ouverte » a priori mais s'est appuyé sur un panel large de sites, à partir de plusieurs répertoires, annuaires, listes etc. (cf. Schöpfel & Prost, 2010). Cette approche impacte nécessairement aussi le 2^{e} volet du projet, l'étude des usages. Mais dans la mesure où seulement très peu de sites ont publié leurs statistiques, cette diversité a moins d'importance.

La deuxième remarque concerne la terminologie et plus généralement, l'utilisation de l'anglais. Pour clarifier certains termes, nous avons ajouté un glossaire, avec quelques définitions, synonymies et traductions. Cependant, ce glossaire est provisoire et fera si possible l'objet d'un travail plus approfondi avec des partenaires anglais, allemands et japonais. Nous avons laissé certaines citations en anglais, sans traduction, afin de faciliter la compréhension des positions et approches à la source.

2. Méthodologie

La méthodologie globale a été décrite dans la première partie du rapport (Schöpfel & Prost, 2010). Pour toutes les archives ouvertes, nous avons cherché des données d'usage en termes d'accès en ligne : une information sur les consultations, visites, téléchargements etc.

La collecte des données en ligne a été faite en équipe, par quatre personnes différentes et avec un contrôle mutuel.

L'étude s'est déroulée en quatre 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.

Afin de tester l'analyse des fichiers log appliquée à une archive ouverte, nous avons ajouté un 2e volet à cette enquête, une étude de cas de l'archive ouverte IRIS[15].

L'étude des statistiques d'IRIS a commencé en octobre 2009 et s'est poursuivie jusqu'en juillet 2010.

L'extraction des statistiques a été effectuée à l'aide d'un accès autorisé aux données via le logiciel Urchin (cf. annexe C) installé sur le serveur d'IRIS.

Le logiciel Urchin a systématiquement filtré et éliminé le trafic généré par des robots. Nous avons également essayé de déterminer la part de l'utilisation interne, c'est-à-dire l'utilisation d'IRIS par les utilisateurs et personnels de la bibliothèque universitaire et du service commun de documentation de Lille 1.

L'étude de cas applique l'approche quantitative de l'équipe CIBER, c'est-à-dire l'analyse des fichiers log en fonction des critères du schéma de 2009 (cf. tableau 1), malgré nos réserves et en vérifiant systématiquement la définition, la faisabilité et l'intérêt de chaque indicateur.

La période d'utilisation analysée a été fixée du 1/1/2009 au 31/12/2009. Les résultats ont été intégrés au fur et à mesure dans un wiki de projet. Ce rapport en contient les principaux éléments.

[15] Archive institutionnelle de l'université Lille 1 (USTL) http://iris.univ-lille1.fr pour la diffusion des thèses (dépôt obligatoire depuis 2008) et documents numérisés en histoire des sciences. La plate-forme pourrait migrer sur l'application ORI-OAI.

3. Résultats de l'enquête

Nous présentons d'abord quelques résultats de l'utilisation des archives ouvertes en France entre 2008 et 2009.

Nous avons trouvé des 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 des chiffres d'utilisation.

Ceci étant, certaines études mélangent « utilisation-dépôt » et « utilisation-accès ». Ainsi, nous avons trouvé deux bilans où sous la rubrique « utilisation » seul le nombre des dépôts avait été compté (Gouat, 2008 et 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.

3.1. Statistiques à destination des auteurs et institutions

Les archives de HAL fournissent des statistiques d'utilisation aux dépositaires et auteurs enregistrés, pour chaque document et pour l'ensemble des documents, avec le format suivant :

1. La durée de la mise en ligne ou l'âge du document (en année, mois, jour).
2. Accès à la fiche concise des métadonnées (consultations).
3. Accès à la fiche étendue (= détaillée) des métadonnées (consultation).
4. 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. L'université de Brest l'a fait pour son archive HAL-UBO en séparant thèses et articles (Bertignac & Gac, 2009).

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 maxima, minima et moyens de téléchargements pour les dépôts de chaque laboratoire[16].

3.2. L'usage par type de document

Nous avons trouvé quatre études qui donnent une information différenciée 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), 99% des articles et 100% des thèses ont été consultés au moins une fois. Le nombre moyen de consultations par article est 49, celui par thèse est de 109 (Malotaux, 2009).

[16] 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.

Pour l'archive institutionnelle de l'université de Bretagne Occidentale à Brest (HAL-UBO), « les thèses sont aujourd'hui les documents (…) les plus consultés » et représentent 17 des 25 documents les plus consultés (Bertignac & Gac, 2009).

Le site de ParisTech[17] contient une liste des « TOP 20 des thèses en ligne » dans PASTEL, en termes de téléchargements 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 et 2009).

Il est intéressant de noter que ces quelques statistiques confirment sans exception l'intérêt de la littérature grise dans ces archives (cf. aussi Schöpfel et al., 2009).

3.3. 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). Une partie des statistiques est communiquée avec explication et interprétation ; d'autres sont mises en ligne d'une façon brute, parfois probablement aussi sans réelle intention de les rendre publiques. En voici un exemple[18] :

Analysed requests from Wed-24-Jun-2009 08:57 to Mon-16-Nov-2009 12:44 (145.16 days).
(Figures in parentheses refer to the 7-day period ending 16-Nov-2009 00:00).
Successful requests: 132,810 (6,975)
Average successful requests per day: 914 (996)
Successful requests for pages: 132,810 (6,975)
Average successful requests for pages per day: 914 (996)
Failed requests: 84 (0)
Distinct files requested: 530 (526)
Distinct hosts served: 40,015 (3,609)
Corrupt logfile lines: 55
Unwanted logfile entries: 2,743,109
Data transferred: 172.81 gigabytes (9.18 gigabytes)
Average data transferred per day: 1.19 gigabytes (1.31 gigabytes)

Figure 2 : Exemple d'un site de fichier logs sur le Web

Les établissements utilisent des outils différents qui ne fournissent pas (toujours) le même type d'information. Entre autre, nous avons identifié Google Analytics / Sitemap, Webalizer Xtended, AWStats[19], PhpMyVisite, Analog. Parmi ces outils, on trouve 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.

[17] http://pastel.paristech.org/apropos/stat.html
[18] Serveur des thèses de l'INP Toulouse, rapport des logs à partir du logiciel "Analog" http://ethesis.inp-toulouse.fr/stats/statsall.html (16 novembre 2009)
[19] AWStats, un logiciel open source et gratuit, est également utilisé par le CLEO pour les statistiques d'utilisation de la plate-forme Revues.org.

Accès sur le site : A partir des adresses IP, ces logiciels donnent une information sur la provenance (pays) des visiteurs[20] ou 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. On peut savoir si un utilisateur est arrivé directement sur le site, via un site référent (lien) ou via un moteur de recherche, et lequel. Les quelques données en ligne sont très claires : la plupart des consultations se font via Google ; l'accès direct ou à partir de sites référents ou d'autres moteurs de recherche (Yahoo etc.) reste marginal[21]. 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).

Comportement sur site : On trouve surtout une information 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). Mais on retrouve aussi d'autres éléments : une analyse des « visites actives » (= sessions avec navigation et/ou recherche), le taux de rebond, le comptage du nombre des pages consultées, la page d'entrée et/ou celle de sortie (= dernière page consultée avant de quitter le site ou de fermer la session). On trouve des chiffres cumulés pour un mois ou une année mais aussi des moyennes journalières ou mensuelles. Parfois, ces mesures sont mises en rapport avec le nombre ou le type des visiteurs.

Il s'agit d'une information riche et détaillée mais très hétérogène, peu exploitée, mal définie. Il est quasiment impossible d'en déduire une conclusion plus générale sur l'utilisation des archives ouvertes en France ou de comparer les sites entre eux.

[20] Pour OATAO entre avril 2008 et février 2009 : 57% France, avec Etats-Unis et Union Européenne 80% (Malotaux, 2009).

[21] 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).

4. Résultats de l’étude de cas IRIS

L’intérêt principal de cette étude de cas est d’appliquer la méthodologie d’une analyse des fichiers log à une archive ouverte. Les résultats chiffrés servent d’illustration de l’approche méthodologique, et nous n’allons pas tenter ici une interprétation par rapport au site en question.

L’archive ouverte IRIS succède à un des premiers sites français dédiés à la littérature grise (GRISEMINE, cf Claerebout, 2003). Elle a la particularité de regrouper à la fois de la littérature grise produite à l’Université de Lille 1 - essentiellement des thèses, HDR et DSR – et un fonds patrimonial en histoire des sciences. A la fin de l’année 2009, IRIS compte 1 036 documents, dont 67,2% sont des documents gris. Au 31 août 2010, IRIS regroupe 1 289 documents.

S’agit-il d’une archive institutionnelle au sens stricte ? La discussion est ouverte ; la page d’accueil du site présente IRIS comme une « bibliothèque numérique ». Le projet de Lille 1 ouvre entre autre des pistes vers un dépôt des ressources pédagogiques numériques, de la littérature grise (documents non publiés) et des documents numérisés (patrimoine scientifique). Si nous appliquons la définition pragmatique de Smith (2008) qui parle d’une « grande variété de documents en format numérique », il n’y a pas de doute qu’IRIS constitue effectivement un exemple particulier de la grande famille des archives institutionnelles.

Nous avons limité l’analyse à l’année 2009. La présentation des résultats suit le schéma cadre des indicateurs de l’équipe CIBER dont les concepts figurent en italique et entre parenthèses dans les sous-titres (cf. tableau 1 et Nicholas et al., 2009b).

4.1. Indicateurs d’activité (*activity metrics*)

4.1.1. Consultations (*number of pages viewed*)

La consultation d’une page correspond à l’affichage d’une page Web dans un navigateur, à la demande de l'explorateur d'un visiteur (internaute) pour une page Web visualisable, généralement un fichier HTML. L’indicateur désigne donc le nombre de fois où une page est affichée ou rendue. CIBER définit la consultation (*page view*) comme “a 'complete' item returned by the server to the client in response to a user action” (Nicholas et al, 2009) ou comme “requests made” (Nicholas et al. 2005). Dans la terminologie COUNTER, tout cela correspondrait au « successful request » (requête réussie).

Le terme de page reste assez ambigu car il recouvre chez Nicholas et al. la liste des numéros d’une revue, le sommaire d’un numéro, les résumés, le texte intégral en HTML ou en PDF mais aussi les schémas, illustrations etc. Pour IRIS, le terme de page concerne aussi bien les interfaces de recherche et de navigation que les notices, les documents, pages d’accueil ou de commentaire etc. Et que veut dire *item* ? Parfois ce terme est utilisé comme synonyme de page (terme générique), parfois il désigne une notice (métadonnées) et/ou toute autre page à l’exception du document en texte intégral. Voici le tableau pour 2009 :

janv-09	29 324
févr-09	29 653
mars-09	34 766
avr-09	27 024
mai-09	27 970
juin-09	25 930
juil-09	16 788
août-09	15 497
sept-09	34 293
oct-09	37 978
nov-09	26 729
déc-09	17 942
Total	323 894

Tableau 4 : IRIS : Nombre de consultations par mois (2009)

En 2009, IRIS a reçu 323 894 consultations en termes de « visualisation de page », l'activité mensuelle allant de 15 000 en août à 38 000 en octobre. A partir de ces données, on pourrait dériver plusieurs indicateurs, dont la moyenne par jour ou mois, la variation par mois (en pourcentage), la moyenne par visiteur unique etc. Afin de faciliter la lisibilité du résultat, on pourrait aussi ajouter une représentation graphique avec une courbe de tendance ou ligne de régression.

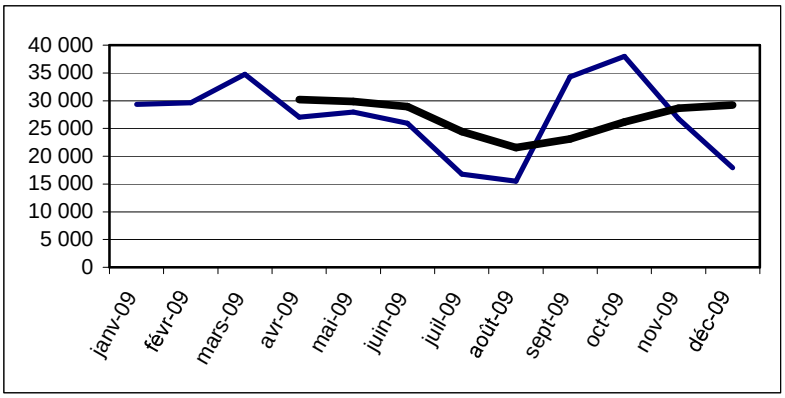

Figure 3 : IRIS : Nombre de consultations par mois, avec courbe de tendance (2009)

Figure 3 propose une courbe de tendance basée sur la moyenne mobile d'une période de quatre mois. Nous rappelons que ces chiffres ne contiennent pas l'activité générée par des robots.

4.1.2. Téléchargements (*number of full-text downloads*)

Le terme "téléchargement" désigne toute opération de transfert d'information du site en question (serveur) vers un poste ou ordinateur distant (client). Sont concernés des images, programmes, documents, données etc.

Le logiciel Urchin classe la popularité de tous les téléchargements en fonction du nombre de succès (demandes) et donne les pourcentages relatifs. Un téléchargement est déterminé par le suffixe de fichier et les variables de configuration Urchin.

Pour IRIS, le logiciel ne compte pas uniquement le téléchargement des documents en texte intégral mais aussi, au moins, des notices. Les chiffres pour 2009 se répartissent ainsi :

janv-09	86 479
févr-09	16 540
mars-09	17 839
avr-09	19 751
mai-09	18 315
juin-09	12 653
juil-09	8 868
août-09	8 072
sept-09	11 021
oct-09	16 331
nov-09	21 429
déc-09	9 530
Total	246 828

Tableau 5 : IRIS : Nombre de téléchargements, version non corrigée (2009)

La courbe suit celle des consultations, à l'exception du mois de janvier qui présente une anomalie flagrante. La vérification des statistiques révèle qu'en janvier 2009, un document particulier a été téléchargé plus de 70 000 fois. Il n'y a aucune explication à cela, sauf celui d'un simple *bug*.[22] Voici le tableau corrigé :

janv-09	16 039
févr-09	16 540
mars-09	17 839
avr-09	19 751
mai-09	18 315
juin-09	12 653
juil-09	8 868
août-09	8 072
sept-09	11 021
oct-09	16 331
nov-09	21 429
déc-09	9 530
Total	176 388

Tableau 6 : IRIS : Nombre de téléchargements, version corrigée (2009)

Dans la suite de notre étude, nous utiliserons ces statistiques corrigées. Voici le graphique des téléchargements, de nouveau avec une courbe de tendance (figure 4) :

[22] L'explication est peu satisfaisante mais c'est la seule explication avancée par le service informatique.

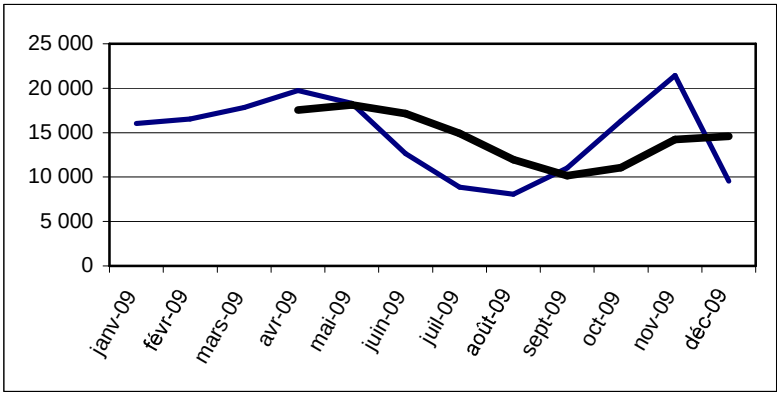

Figure 4 : IRIS : Nombre de téléchargements, avec courbe de tendance (2009)

4.1.3. Sessions (*number of sessions conducted*)

Une session ou visite[23] est définie comme une série de pages Web consultées de façon consécutive durant un laps de temps défini. Une session est commencée lorsque le visiteur entre sur le site et se termine soit lorsqu'il quitte le site ou au moment où l'explorateur est fermé, soit après une certaine période d'inactivité.

CIBER place les *search sessions* au cœur de l'analyse des fichiers log, au même titre que les consultations (Nicholas et al., 2005 et 2007). Jana & Chatterwee (2004) considèrent le nombre de sessions comme la meilleure mesure pour le nombre de visiteurs uniques. Voici les statistiques pour IRIS :

janv-09	2 166
févr-09	2 311
mars-09	2 907
avr-09	2 625
mai-09	2 883
juin-09	2 528
juil-09	1 904
août-09	1 704
sept-09	2 352
oct-09	2 745
nov-09	2 803
déc-09	1 643
Total	28 571

Tableau 7 : IRIS : Nombre de sessions (2009)

[23] Visite et session sont souvent utilisées comme synonymes. Pourtant, il existe une distinction : "A visit is defined as a series of page requests from the same uniquely identified client with a time of no more than 30 minutes between each page request. A session is defined as a series of page requests from the same uniquely identified client with a time of no more than 30 minutes and no requests for pages from other domains intervening between page requests. In other words, a session ends when someone goes to another site, or 30 minutes elapse between pageviews, whichever comes first. A visit ends only after a 30 minute time delay. If someone leaves a site, then returns within 30 minutes, this will count as one visit but two sessions. In practise, most systems ignore sessions and many analysts use both terms for visits. Because time between pageviews is critical to the definition of visits and sessions, a single one pageview event does not constitute a visit or a session (it is a "bounce")."
http://en.wikipedia.org/wiki/Web_analytics (consulté le 15 juillet 2010)

De nouveau, ces chiffres suivent assez bien les autres courbes d'activité (cf.figure5).

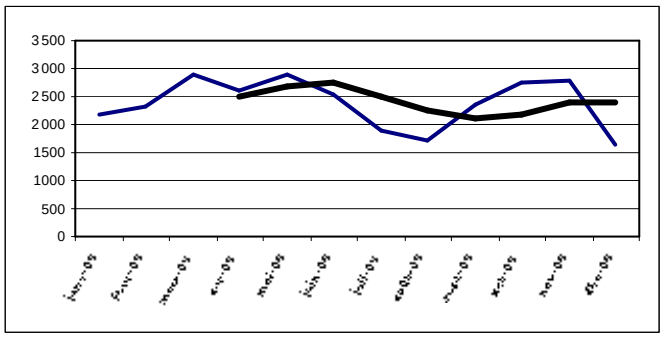

Figure 5 : IRIS : Nombre de sessions, avec courbe de tendance (2009)

A partir de ces chiffres, on peut calculer le nombre moyen par jour (70-100), et on peut (comme CIBER) mettre les sessions en rapport avec l'origine du trafic (international, origine inconnue, du pays) et les services d'accès (moteur de recherche, robots...). Peut-on pour autant conclure qu'IRIS reçoit la visite d'environ 2 000 à 3 000 internautes par mois ? Il faudrait probablement rester prudent et mettre ce chiffre en relation avec les adresses IP identifiées (cf. 4.3.).

4.1.4. Profondeur de visite (*site penetration*)

Le terme de profondeur de visite désigne le nombre moyen d'accès aux pages, par visite. Ou plus simplement, le nombre de pages consultées lors d'une session.

Nicholas et al. (2005) interprètent cette donnée comme une mesure centrale de l'empreinte numérique ("digital fingerprint") et de l'activité sur un site : "Another, more powerful, way of examining the number of items viewed is to categorise user search sessions by the number of items viewed – we call such an analysis "site penetration". Site penetration is a deep log technique which takes log data much further than is the norm, and offers an extremely good platform for characterising the information-seeking behaviour of sub-groups of users."

Nicholas et al. (2005) regroupent les sessions en quatre catégories qu'ils utilisent ensuite selon le type du document, les internautes etc. Nous avons appliqué les mêmes catégories pour IRIS. Voici le résultat :

Consultations	Nb de sessions	En %
1-3 pages par session	13 768	48%
4-10 pages par session	7 531	26%
11-19 pages par session	3 319	12%
20+ pages par session	4 053	14%
	28 571	

Tableau 8 : IRIS : Profondeur de visite (2009)

Dans 48% des sessions, l'internaute ne consulte que 1-3 pages avant de répartir. Mais dans 14% des sessions, 20 pages ou plus sont visualisées. Comment expliquer ce chiffre ? En fait, il s'agit surtout d'utilisation interne (cf. 5.5.).

Afin d'obtenir un indicateur approximatif, nous avons également divisé pour chaque mois le nombre de consultations par le nombre des sessions. Mais cette procédure pose deux problèmes. D'une part, le mode de calcul (moyenne arithmétique) n'est pas adapté, du fait de la distribution non normalisée. Et d'autre part, Urchin ne détaille pas au-dessus de 20 consultations ce qui biaise tout calcul de moyenne.

4.1.5. Durée de la consultation d'une page (*time spent viewing a page*)

Combien de temps le visiteur reste-t-il sur une page ? Nicholas et al. (2005) font le lien entre cette donnée et la typologie des pages visitées, en calculant le médian pour chaque cas de figure :

1. Liste des numéros : 7 sec
2. Table des matières :10 sec
3. Résumé : 1 sec
4. Texte intégral HTML : 21 sec
5. Texte intégral PDF : 18 sec

Le logiciel Urchin produit une liste de la durée moyenne d'utilisation pour chaque page hormis les pages de sortie. La durée moyenne de consultation de la page est indiquée en heures:minutes:secondes. Voici les statistiques pour IRIS, pour quatre catégories de pages :

	Durée totale	Nb pages	Nb visites	Durée visite
Notice, interface (handle)	42:18:16	1414	121 851	1 sec
Document (bitstream)	04:07:13	208	911	16 sec
Requête OAI	00:35:49	1	559	4 sec
Votre commentaire	00:28:50	1	1274	1 sec

Tableau 9 : IRIS : Durée moyenne de consultation d'une page en secondes (2009)

Les notices et les interfaces attirent le plus de trafic, et la durée cumulée dépasse les 40 heures – environ 85% de la durée totale des sessions. Mais consulter les notices etc. ne prend pas beaucoup de temps, en moyenne autour d'une seconde. Tandis que l'accès au document en texte intégral dure en moyenne bien plus longtemps (16 secondes). Cette durée d'accès au document correspond assez bien aux résultats moyens de CIBER (18-21 secondes).

4.1.6. Durée d'une session (*time spent on a session*)

COUNTER définit ce terme comme « duration », ou la durée moyenne d'une session. Le logiciel Urchin mesure la différence entre le moment du premier accès et le dernier accès pages d'une session. Les sessions ne comportant qu'un seul accès pages sont considérées comme ayant une durée de 0-10 secondes. Il est à noter que les visiteurs peuvent passer plus de temps au dernier accès page d'une session, mais que seul le moment de chargement est enregistré. Les statistiques pour IRIS :

Durée	Nb de sessions	% des sessions
0-10 sec	14 031	49%
11-30 sec	2 467	9%
31-60 sec	2 038	7%
1-3 min	3 482	12%
3-10 min	3 349	12%
10-30 min	2 292	8%
30+ min	912	3%
Total	28 571	

Tableau 10 : IRIS : Durée moyenne d'une session (2009)

La moyenne médiane de la durée d'une visite se situe autour de 10 secondes. Cependant, on constate aussi qu'il y a des sessions qui durent bien plus longtemps. Environ 10% des sessions durent au moins 10 minutes – un temps assez long pour un site Web.

4.1.7. Nombre de recherches par session (*number of searches undertaken in session*)

Il s'agit du nombre de recherches effectuées au cours d'une seule session. Nous avons calculé ici la moyenne du nombre de pages visualisées pendant une session. La même valeur se trouve dans le résumé du trafic. Voici les statistiques pour IRIS :

janv-09	14
févr-09	13
mars-09	12
avr-09	10
mai-09	10
juin-09	10
juil-09	9
août-09	9
sept-09	15
oct-09	14
nov-09	10
déc-09	11

Tableau 11 : IRIS : Nombre de recherches par session (2009)

D'après ces statistiques, une session compte en moyenne 9-15 pages. Mais pour les raisons indiquées plus haut (cf. 4.1.4), cet indicateur ne paraît ni satisfaisant ni fiable. En plus, quand on filtre l'utilisation du site par l'équipe du SCD de Lille 1, les statistiques révèlent une différence significative entre l'usage interne (jusqu'à 30 pages/recherches) et externe (au-dessous de 10 pages par session).

4.1.8. Nombre de documents consultés (*number of sources used*)

La définition du terme est relativement simple : il s'agit du nombre de documents déposés dans l'archive et utilisés au moins une fois pendant la période en question. En limitant l'extraction des pages par Urchin aux seules pages identifiées par "bitstream" qui correspondent à des fichiers PDF, on obtient 319 documents consultés en 2009, dont plusieurs

avec des adresses apparemment invalides[24]. Cela correspond à 30,8% du contenu d'IRIS en 2009.

20 documents ont été consultés cinq fois ou plus. Mais d'après Urchin, plus de 700 documents n'ont jamais été visualisés. Ce pourcentage d'environ 70% de documents non utilisés est étonnant et au fond, incompréhensible.

4.1.9. Nombre de consultations par document (*number of views per source*)

Voici le début de la liste des documents d'IRIS, avec les 20 premiers identifiants « bitstream » de DSpace correspondant à des fichiers PDF :

/dspace/bitstream/	Invalide	132	14,49%
/dspace/bitstream	Invalide	63	6,92%
/dspace/bitstream/1908/	Invalide	54	5,93%
/dspace/bitstream/1908	Invalide	30	3,29%
/dspace/bitstream/1908/171/1/5037	Thèse	25	2,74%
/dspace/bitstream/1908/1044/	Invalide	22	2,41%
/dspace/bitstream/1908/995/1/50374-200...	Document	21	2,31%
/dspace/bitstream/1908/223/1/	Thèse	12	1,32%
/dspace/bitstream/1908/281/1/	Thèse	10	1,10%
/dspace/bitstream/1908/223/	Invalide	8	0,88%
/dspace/bitstream/1908/255/1/	Thèse	7	0,77%
/dspace/bitstream/1908/281/1	Thèse	7	0,77%
/dspace/bitstream/1908/241/1/	Thèse	7	0,77%
/dspace/bitstream/1908/281/	Thèse	7	0,77%
/dspace/bitstream/1908/993/1/	Invalide	7	0,77%
/dspace/bitstream/1908/795/1/pas de fichier	Invalide	7	0,77%
/dspace/bitstream/1908/300/1/	Thèse	6	0,66%
/dspace/bitstream/1908/992/1/	Thèse	6	0,66%
/dspace/bitstream/1908/229/	Invalide	5	0,55%
/dspace/bitstream/1908/265/1/	Thèse	5	0,55%

Tableau 12 : IRIS : Nombre de consultations par document – les 20 premiers identifiants (2009)

Comme indiqué plus haut, les quatre premières adresses renvoient sur des pages invalides ; parmi les 20 références, neuf renvoient sur des pages invalides. Ceci n'est pas la seule anomalie du tableau. Par exemple, on y trouve des doublons. Mais sans lien vers les métadonnées, un contrôle paraît bien fastidieux. Nous ne l'avons pas entrepris.

[24] C'est-à-dire, changement ou suppression d'adresse entre la consultation en 2009 et l'analyse des statistiques en 2010.

4.2. Indicateurs de la recherche d'information (*information seeking characteristics*)

4.2.1. Documents téléchargés par session (*number of sources used in a session*)

Nicholas et al. (2008) publient le nombre de revues consultées en une session - combien de consultations concernent un seul titre, combien deux titres ou plus ? L'explication ou la portée de cette analyse semble assez limitée. D'autres données, comparables à celle-ci sont le nombre moyen des pages consultées pendant une session ou le nombre moyen des articles consultés pendant une session (Nicholas et al, 2009).

Comme Urchin ne permet pas de calculer cet indicateur, nous avons calculé le nombre de téléchargements par session (corrigé pour janvier 2009). Voici le tableau :

janv-09	7
févr-09	7
mars-09	6
avr-09	8
mai-09	6
juin-09	5
juil-09	5
août-09	5
sept-09	5
oct-09	6
nov-09	8
déc-09	6

Tableau 13 : IRIS : Nombre de pages téléchargées par session (2009)

D'après ces chiffres, les visiteurs d'IRIS ont téléchargé en moyenne 5-8 pages par session. Mais cela n'indique pas l'utilisation réelle des documents déposés dans IRIS.

4.2.2. Documents consultés (*names of sources used and not used*)

Cette donnée contient la description bibliographique des documents consultés au moins une fois, et de tous ceux qui n'ont pas été consultés. La description bibliographique se limite au titre de la revue consultée, généré à partir de l'ISSN (Nicholas et al., 2007). L'analyse se limite à une sorte de « hit-parade » des *top 20 journals used*.

Notre propre analyse se contente donc d'une extraction (manuelle) du titre et de l'auteur des dix documents les plus consultés en 2009 (métadonnées) :

	Auteur	Titre	Type	Année	Thématique
1.	Vittecoq, E	Du crash-test aux essais mono-filamentaires, quelques apports dans le domaine de la caractérisation expérimentale du comportement de matériaux et de structures	HDR	2004	Mécaniques de l'ingénieur
2.	Bourdet, D	Les pratiques communicationnelles médiatisées des étudiants roumains à Iasi	Thèse	2005	Sociologie
3.	Bouchart, V	Étude expérimentale et modélisation micromécanique du comportement et de l'endommagement des élastomères renforcés	Thèse	2007	Mécaniques de l'ingénieur
4.	Hennebelle, T	Investigation chimique, chimiotaxonomique et pharmacologique de Lamiales productrices d'antioxydants : Marrubium peregrinum, Ballota larendana, Ballota pseudodictamnus (Lamiacées) et Lippia alba (Verbénacées)	Thèse	2006	Médecine, Pharmacologie
5.	Pb d'accès				
6.					
7.	Andoulsi, R	Étude d'une classe de systèmes photovoltaïques par une approche bond graph : modélisation, analyse et commande	Thèse	2001	Ingénierie
8.	Pb d'accès				
9.	Kestelyn, X	Modélisation vectorielle multimachines pour la commande des ensembles convertisseurs machines polyphasés	Thèse	2003	Electrotechnique
10.	Solarski, Sl	Développement de nouveaux filaments de polylactide nanocomposites	Thèse	2006	Chimie organique

Tableau 14 : IRIS : Les dix premiers documents les plus consultés (2009)

Le premier document est celui victime d'un bug (cf. plus haut). Trois documents de cette liste ne sont plus accessibles sous leur adresse initiale ; sans correspondance avec les métadonnées, il n'est pas facile de les identifier.

Consultation ne veut pas dire téléchargement. A titre d'exemple, le document le plus téléchargé en 2009 est une thèse de 2005, « Les microstructurations dans les fibres optiques » d'Emmanuel Kerrinckx, qui ne figure pas parmi les documents les plus consultés.

De même, on ne trouve aucun document à caractère patrimonial (histoire des sciences) sur cette liste. Manque d'intérêt ? Problème d'indexation, d'accès ou de visibilité ?

4.2.3. Année de publication (*age of source used*)

Le projet COUNTER définit l'âge des documents consultés comme l'année au cours de laquelle un article, item, fascicule ou volume est publié la première fois sur quelque support que ce soit. La terminologie de CIBER est plus large : "publication year/age of article item viewed" (Nicholas et al, 2008) ou "median age of article viewed (in months)" (Nicholas et al, 2009).

Avec Urchin, nous avons analysé l'âge des 100 documents les plus téléchargés. Ces documents correspondent à 56% du total des téléchargements en 2009 (tableau 15) :

Année	Nb docs	Téléchargements
<2000	2	3 427
2000	6	3 689
2001	13	11 952
2002	9	7 661
2003	20	16 598
2004	11	9 237
2005	13	16 255
2006	18	19 466
2007	5	8 713
2008	3	2 229
2009	0	0
2010	0	0
	100	92 227

Tableau 15 : IRIS : Année de publication des 100 documents les plus téléchargés (2009)

Ces statistiques ont été produites à partir du nom de fichier qui contient, au moins pour les travaux universitaires, l'année de publication (soutenance pour les thèses). Ceci n'est pas précis ni exhaustif. Notamment pour les documents à caractère patrimonial, le nom des fichiers suit une autre syntaxe. Une telle mesure reste donc provisoire, sans lien avec les données descriptives des documents.

Nous n'avons pas d'éléments d'explication pour la baisse des chiffres pour les thèses plus récentes, soutenues en 2007 et 2008 – l'effet d'un dépôt décalé ?

Nicholas et al. (2008) présentent l'âge des documents par catégorie. Ils proposent trois ou quatre catégories : la période en cours un a deux ans (*current*), les cinq dernières années (*declining*), les vieux documents, avec plus de cinq ans d'âge (*old*). Dans une autre étude (Nicholas et al., 2007), ils utilisent quatre catégories : l'année en cours, un à trois ans, quatre à sept ans, plus de sept ans. Voici les mêmes statistiques regroupées en trois catégories :

2008-2009	3	2 229
2003-2007	67	63 269
<2003	30	26 729

Tableau 16 : IRIS : Age de publication des 100 documents les plus téléchargés (2009)

Cet indicateur peut paraître logique sur le site d'un éditeur qui met en ligne des articles originaux. Cependant, par rapport à une archive ouverte, il y a un 2^{e} critère, la date du dépôt et/ou de la mise en ligne. Là encore, sans accès aux métadonnées il est impossible d'effectuer cette analyse.

4.2.4. Taille des fichiers téléchargés (*size of source used*)

La taille des fichiers téléchargés est mesurée en octets transférés du serveur au client. Il s'agit d'un indicateur pour l'activité sur le site. En additionnant la taille des fichiers demandés par les visiteurs du site, on obtient une évaluation globale du volume du trafic. Voici le tableau pour 2009 :

pdf	334,65 GB	98,24%
(no type)	3,19 GB	0,94%
autres	2,79 GB	0,82%
Total	340,63 GB	100,00%

Tableau 17 : IRIS : Taille des fichiers téléchargés, en gigabytes (2009)

Sans surprise, pratiquement tous les fichiers téléchargés sont en PDF, le format de dépôt des documents d'IRIS. Plus surprenante est la diversité de format quand on regarde les autres fichiers - des formats image (gif ou jpg) mais aussi du HTML ou des formats textes. Mais tous au-dessous du seuil de 0,01% du volume global, donc négligeable

Le 2e tableau contient les mêmes chiffres mais répartis par mois :

janv-09	31,11 GB
févr-09	30,96 GB
mars-09	33,24 GB
avr-09	32,04 GB
mai-09	30,06 GB
juin-09	25,32 GB
juil-09	18,41 GB
août-09	16,86 GB
sept-09	28,45 GB
oct-09	39,16 GB
nov-09	32,60 GB
déc-09	22,44 GB
	340,65 GB

Tableau 18 : IRIS : Taille des fichiers téléchargés, en gigabytes par mois (2009)

Cet indicateur suit visiblement assez fidèlement l'activité en termes de trafic (consultations, téléchargements). Il est certain aussi par contre qu'il dépend directement du format, du contenu et de la qualité des fichiers. D'après ces statistiques, la taille moyenne d'un fichier téléchargé se situe autour de 2 Mo ce qui correspondrait globalement au poids d'une thèse en format PDF dans IRIS.

4.2.5. Approche de recherche (*search approach adopted*)

Comment les visiteurs se comportent-ils sur le site ? Quels sont les chemins de navigation les plus empruntés ? L'analyse tente d'extraire et d'interpréter une « combination of items » en termes de stratégie de recherche (Nicholas et al., 2005). Dans son étude des revues de Blackwell, CIBER sépare article, résumé et sommaire du fascicule ce qui donne plusieurs stratégies possibles :

1. Seulement sommaire
2. Résumé avec sommaire
3. Résumé sans sommaire
4. Article/résumé avec sommaire
5. Article/résumé sans sommaire

6. Article avec sommaire
7. Article sans sommaire

Parmi ces stratégies, 1 (seulement sommaire) et 3 (seulement résumé) correspondent chacune à environ 40% des stratégies sur site.

Nicholas et al. (2007) distinguent plusieurs modes de recherche (simple, avancé, expert). Cette distinction s'appuie sur le portail d'OhioLINK et sur l'utilisation de ses interfaces de recherche, mais pas sur une analyse de la stratégie appliquée (suite des pages etc.). Il faudra donc "ré-inventer" le contenu de cette donnée en fonction de la qualité des fichiers log.

Le logiciel Urchin ne conserve, par défaut, que trois pages pour caractériser une navigation. En 2009, 5 392 de ces « navigations » ont été répertoriées. Voici les dix approches les plus importantes :

1. /dspace/ /dspace/styles.css.jsp	Page d'accueil Feuille de style	2 400	8,40%
2. /dspace/ /dspace/styles.css.jsp /dspace/handle/1908/3	Page d'accueil Feuille de style Catégorie des thèses	1 602	5,61%
3. /dspace/ /dspace/styles.css.jsp /dspace/simple-search	Page d'accueil Feuille de style Résultat de la recherche	1 472	5,15%
4. /dspace/ /dspace/styles.css.jsp /dspace/handle/1908/54	Page d'accueil Feuille de style Classement par discipline	1 175	4,11%
5. /dspace/	Page d'accueil	511	1,79%
6. /dspace/ /dspace/styles.css.jsp /dspace/	Page d'accueil Feuille de style Page d'accueil	434	1,52%
7. /dspace/dspace/handle/1908/156 /dspace/styles.css.jsp	Liste des thèses Feuille de style	418	1,46%
8. /dspace/handle/1908/3 /dspace/styles.css.jsp /dspace/handle/1908/3	Catégorie des thèses Feuille de style Catégorie des thèses	415	1,45%
9. /dspace/ /dspace/styles.css.jsp /dspace/handle/1908/31	Page d'accueil Feuille de style Identifiant invalide	386	1,35%
10. /dspace/styles.css.jsp	Feuille de style	358	1,25%

Tableau 19 : IRIS : Dix modes de navigation, avec le nombre de sessions (2009).

Ces données brutes sont difficiles à exploiter ou à interpréter ; comment notamment expliquer la présence de la feuille de styles. Quand on filtre cette feuille de style, il reste des navigations assez particulières, des doubles-clics, des accès à une seule page, mais aussi des stratégies un peu plus élaborées comme par exemple :

Page d'accueil -> interface de recherche avancée -> plan de classement
Interface de recherche avancée -> liste des thèses en économie

Par où les internautes entrent-ils sur le site ? Voici les pages d'entrée préférées :

Page d'accueil	10 176	35,62%
Thèses de l'USTL par collection	1 049	3,67%
Thèses de l'USTL interface recherche	1 048	3,67%
Votre commentaire	602	2,11%
Feuille de style	559	1,96%
Thèses en physique	421	1,47%

Tableau 20 : IRIS : Pages d'entrée préférées, avec nombre de sessions (2009)

Un tiers des sessions commence sur la page d'accueil d'IRIS. D'autres visites débutent directement sur une liste de thèses, une interface de recherche ou la feuille de style. Cela s'explique plus facilement, par exemple par le référencement et le ranking des pages d'IRIS dans Google qui renvoient une recherche « IRIS USTL » sur la page *Thèses de l'USTL*[25].

4.2.6. Nombre de termes par requête (*number of search terms used in search*)

Combien de mots les internautes entrent-ils dans les moteurs de recherche pour trouver le site IRIS ? Et quels mots ou phrases utilisent-ils ?

Sous la notion des « meilleures clés de recherche », Urchin produit une liste des phrases/mots clés réels, avec l'affichage des entrées en ordre, selon le nombre de visiteurs correspondant à chacune d'elles. Voici un extrait de cette liste, avec les dix entrées les plus utilisées :

Termes	Nb de recherches	%
Iris	1 244	9,48%
Iris+lille+1	105	0,80%
E+classement	73	0,56%
these+physique	64	0,49%
thèse+physique	64	0,49%
Iris+	58	0,44%
travail+de+fin+d'études	54	0,41%
Iris+lille1	50	0,38%
Iris+lille	40	0,30%
travail+de+fin+d'étude	39	0,30%

Tableau 21 : IRIS : Les dix premières clés de recherche (2009)

9 278 formules différentes de recherche ont été enregistrées en 2009. Les stratégies d'approche sont assez variées, même si on tient compte des différentes variantes d'écriture (avec ou sans accent, singulier ou pluriel etc.). On trouve à la 11e position la recherche « histoire+de+la+physique ». En regroupant les cent premières entrées correspondant à 32% des recherches, on obtient le tableau suivant :

[25] https://iris.univ-lille1.fr/dspace/handle/1908/3

Nombre de termes	Nombre de recherches	%
1	1 416	9
2	791	48
3	446	27
4	210	9
5	88	6
6	9	1
	2 960	100

Tableau 22 : IRIS : Répartition des 100 1ères clés de recherche selon le nombre de termes (2009)

84% des recherches se contentent d'un, de deux ou de trois mots. Un résultat quelque peu banal : c'est le reflet du comportement de la « génération Google ».

4.2.7. Mode de navigation (*form of navigation*)

Que font les internautes une fois arrivés sur le site ? Est-ce qu'ils feuillètent, est-ce qu'ils cherchent et si oui, de quelle façon ? Voici de nouveau quelques éléments à partir d'une synthèse des fichiers log (= 20 967 actions sur le site en 2009) :

Mode de navigation	Nombre	%
Feuilletage	12 544	60%
Recherche	7 761	37%
Authentification	589	2,80%
Autres	73	0,30%
Total	20 967	100%

Tableau 23 : IRIS : Répartition des différents modes de navigation (2009)

Le *browsing* ou feuilletage à partir des listes de sujets, auteurs ou documents semble être le mode de navigation préféré. Ici, l'approche par auteur domine, suivi par la thématique des thèses :

Feuilletage auteur	4 464	58%
Feuilletage sujet thèse	2 275	29%
Feuilletage domaine	444	6%
Feuilletage type de doc	553	7%
	7736	100%

Tableau 24 : IRIS : Les différents modes de feuilletage (2009)

D'après les statistiques, pratiquement toutes les recherches ont été lancées à partir d'une interface de recherche simplifiée :

Recherche simplifiée	12 404	99%
Recherche avancée	140	1%

Tableau 25 : IRIS : Répartition des différents modes de recherche (2009)

La plus grande partie des recherches se fait à partir des interfaces simplifiées, surtout à partir du formulaire sur la page d'accueil (40% des visites).

4.2.8. Chemin d'accès (*from where users arrive from*)

Par quel chemin les internautes sont-ils arrivés sur le site d'IRIS ? D'après les explications de CIBER, cette information tente d'éclaircir "where they were and what they might have been doing prior to arriving at (...)" (Nicholas et al, 2009). En fonction des éléments fournis par les éditeurs, CIBER utilise plusieurs catégories pour ces « referral links » dont les moteurs de recherche, des liens à partir du site de la bibliothèque, l'interface de recherche sur le site en question, le feuilletage (browsing) sur le même site etc.

Voici un tableau synthétique des chemins qui mènent vers IRIS, établi à partir des 100 sites de référence les plus importants, correspondant à 98,82% du trafic.

Google	12 491	44%
Non Google	15 742	56%
dont accès direct	*7 476*	*26%*
dont Lille 1	*6 271*	*22%*
Total	28 233	

Tableau 26 : IRIS : Les chemins d'accès (2009)

44% du trafic arrivent via Google (Google France, Algérie... y compris Google Scholar). 22% à partir de plusieurs sites de Lille 1 (page d'accueil, SCD, service d'authentification), 26% en accès direct. Voici un tableau limité aux seuls moteurs de recherche (et à 15 lignes) ; l'importance de Google saute aux yeux :

www.google.fr	8 614	65,63%
www.google.com	1 478	11,26%
scholar.google.fr	624	4,75%
www.google.dz	394	3,00%
google.co.ma	319	2,43%
Scholar.google.com	283	2,16%
www.google.be	178	1,36%
www.google.ca	150	1,14%
www.bing.com	133	1,01%
Search.yahoo.com	81	0,62%
85.229.132	72	0,55%
www.google.ch	60	0,46%
recherche.aol.fr	57	0,43%
search.live.com	48	0,37%
scholar.google.ca	44	0,34%

Tableau 27 : IRIS : Accès via les moteurs de recherche (2009)

L'analyse des fichiers log permet également de détailler la navigation à partir du site de Lille 1. Voici quelques exemples :

www.univ-lille1.fr/bustl/frame_grisemine_haut.html	4 094	65,02%
www.univ-lille1.fr/bustl/index.php	932	14,80%
http://doc.univ-lille1.fr/Ressources_Electroniques/Theses_electroniques/	586	9,31%
https://sso-cas.univ-lille1.fr/login	240	3,81%
http://doc.univ-lille1.fr/Ressources_Electroniques/Bibliotheque_numerique_en_histoire_des_Sciences/	155	2,46%
http://ustl1.univ-lille1.fr/projetUstl/Recherche/ecolesDoct/AffDoct.htm	103	1,64%
https://sso-cas.univ-lille1.fr/	42	0,67%
www.univ-lille1.fr/bustl/	26	0,41%
http://ustl1.univ-lille1.fr/projetUstl/Recherche/ecolesDoct/affdoct.htm	25	0,40%
http://cridoc.univ-lille1.fr/bustl/grisemineMK/accueil.asp	21	0,33%

Tableau 28 : IRIS : Accès via le site institutionnel de Lille 1 (2009)

La synthèse est assez surprenante même si l'accès à partir du site institutionnel ne concerne qu'un pourcentage limité de l'ensemble (22%) :

Ancien site GRISEMINE : 66%
Site de la BU, plusieurs pages (thèses électroniques, bibliothèque numérique…) : 27%
Service d'authentification Lille 1 (ENT) : 5%
Site institutionnel Lille 1 : 2%

Apparemment, GRISEMINE sert toujours de porte d'entrée d'IRIS, malgré sa fermeture. Il est par contre difficile d'estimer combien d'utilisateurs arrivent via le portail documentaire.

4.3. Information sur les utilisateurs (*user characteristics*)

4.3.1. Domaine scientifique (*subject, discipline*)

Cette donnée est censée caractériser l'utilisateur. Quand on compare les études des fichiers log, on constate qu'en réalité elle recouvre au moins trois niveaux différents : la discipline du chercheur-visiteur du site, le domaine scientifique de son établissement d'affiliation, et la thématique des ressources consultées.

Une autre ambiguïté s'ajoute à ce flou conceptuel : en fait, qu'est-ce qu'un utilisateur, visiteur ou internaute ? CIBER définit le « user » comme une combinaison d'adresse IP et d'informations issues du navigateur (browser) du serveur client (Nicholas et al., 2005). Parfois il s'agit d'une machine (robot), parfois d'un individu ou d'un groupe.

Certains éléments proviennent de l'identification d'utilisateurs autorisés et enregistrés dans un « fichier de lecteurs » maintenu par une bibliothèque, un consortium ou un éditeur.

Dans le cas d'IRIS, il n'existe pas de fichier d'authentification ou de fichier d'information personnelle. Il faut donc se contenter d'une analyse sommaire à l'appui de quelques éléments des fichiers log. A partir des 18 640 adresses IP recensées, le logiciel Urchin propose une extraction des « meilleures adresses IP » avec une arborescence des IP et aussi, celle des « meilleurs domaines ». Voici un tableau :

Domaine	Nombre de sessions	%
Lille 1 STM	1 879	6,60%
CEA Physique nucléaire	177	0,60%
Valenciennes PPD	164	0,60%
Rennes STM	106	0,40%
INSA Lyon STM	84	0,30%
Lyon 1 STM	79	0,30%
Strasbourg GPD	75	0,30%
	28 571	

Tableau 29 : IRIS : Domaines d'appartenance des utilisateurs (2009)

Nous avons mis cette information en relation avec la nomenclature du Ministère de l'Enseignement Supérieur (ASIBU)[26]. Cela donne une première idée d'où viennent les internautes, même si les catégories ASIBU sont très globales, peu précises. Mais est-ce vraiment important ? Tout cela reste marginal (<10% des sessions), trop général, partiel. Et puis, il y a risque de confusion avec l'organisme-employeur de l'utilisateur. Pour cette raison, nous ne sommes pas allés plus loin

4.3.2. Origine géographique de l'utilisateur (*geographical location*)

L'analyse de CIBER s'appuie sur les fichiers commerciaux des éditeurs, c'est-à-dire sur les données personnelles et protégées des utilisateurs autorisés d'accès au service en question (Nicholas et al., 2005).

En absence d'une telle source d'information, il faut se replier sur une approche alternative, sur l'étude de l'origine du trafic (par session) identifiée à partir du domaine.

Le logiciel Urchin produit une liste des principaux domaines avec le nombre de session. Voici le tableau pour 2009 avec les 20 domaines les plus importants :

[26] STM sciences techniques médecine ; GPD grand pluridisciplinaire ; PPD petit pluridisciplinaire

1.	Fr (France)	9 466	33,13%
2.	(no entry)	9 106	31,87%
3.	net (Network)	5 801	20,30%
4.	ma (Morocco)	1 002	3,51%
5.	com (Commercial)	878	3,07%
6.	ca (Canada)	415	1,45%
7.	be (Belgium)	398	1,39%
8.	private-172-net	197	0,69%
9.	ch (Switzerland)	149	0,52%
10.	org (Non-Profit Organizations)	121	0,42%
11.	tn (Tunisia)	101	0,35%
12.	de (Germany)	75	0,26%
13.	ro (Romania)	69	0,24%
14.	It (Italy)	56	0,20%
15.	pl (Poland)	45	0,16%
16.	dz (Algeria)	40	0,14%
17.	es (Spain)	36	0,13%
18.	jp (Japan)	34	0,12%
19.	uk (United Kingdom)	31	0,11%
20.	ci (Cote D'Ivoire)	28	0,1

Tableau 30 : IRIS : Origine géographique des visiteurs (2009)

Comme la majeure partie du trafic Web provient des domaines .net, .com et .org, (principalement basés aux États-Unis), ils sont inclus. Le restant représente surtout des codes de pays qui sont dans la plupart des cas des indicateurs relativement fiables de l'origine du trafic (bien que certains vendent des domaines à n'importe qui). Le terme *no entry* – un tiers des sessions - inclut toutes les adresses IP non résolues.

En regroupant les 12 415 sessions qui peuvent être localisées (= 43%) sans compter les domaines .net, .com et .org, on obtient un tableau plus synthétique par continent :

Europe	10 506	85%
Afrique	1 293	10%
Amérique du Nord	422	3%
Asie/Océanie	117	1%
Amérique latine	73	1%
International	4	0%

Tableau 31 : IRIS : Origine géographique des 12 415 sessions localisées[27] (2009)

[27] En ajoutant les .net, .org et .com comme adresses américaines, cela donnerait ceci :

Europe	10506	55%
Afrique	1293	7%
Amérique du Nord	7222	38%
Asie/Océanie	117	1%
Amérique latine	73	0%
International	4	0%

La réalité devrait se situer quelque part entre les deux tableaux.

La France représente 9466 sessions (= 76%) ; en tout, 11 781 sessions (= 95%) des sessions localisées proviennent des pays francophones.

4.3.3. Organisme employeur de l'utilisateur (*name of organization*)

Nous avons déjà abordé la question de l'identification. Il faut prendre ce concept au sens large – structure d'affiliation, établissement de formation etc. Une exploitation non exhaustive et prudente des domaines des navigateurs enregistrés en 2009 indique 3253 adresses pour les universités (11,4%) et 351 adresses pour les écoles et l'INSA (1,2%). De nouveau, c'est assez marginal, imprécis et peu fiable.

5. Discussion

Notre étude sur le terrain des archives ouvertes partait de deux hypothèses :

1. Il y a davantage d'information sur l'utilisation de ces archives, en termes d'accès (visites) et de téléchargements.
2. Cette information n'est ni exhaustive, ni normalisée.

Nous discuterons d'abord les résultats par rapport à ces deux hypothèses, avant d'aborder l'analyse des fichiers log et l'étude de cas IRIS.

5.1. La dissémination et l'environnement des statistiques d'utilisation

Nous avons effectivement trouvé davantage d'information sur l'utilisation des archives ouvertes qu'en 2008. Plusieurs organismes commencent à informer sur l'usage de leur archive ouverte, sur leur site, sous forme de rapport d'activité ou à travers d'une communication. Parfois on trouve tout simplement des traces de données sur Internet, un peu par hasard sans doute. Tout cela reste néanmoins limité, anecdotique, sans représentativité.

La 2e hypothèse est également confirmée, au moins partiellement : les statistiques d'utilisation de 2009 sont toujours aussi partielles, aléatoires et hétérogènes qu'en 2008 ; il est pratiquement impossible d'effectuer des comparaisons ou d'autres analyses plus poussées.

Il est par contre possible que notre méthodologie - la recherche de statistiques sur le Web – soit passée à côté d'autres documents et données appartenant à la littérature grise ou se trouvant dans les profondeurs du Web.

Plus généralement, la communication sur l'usage se développe dans un environnement favorable. Plusieurs initiatives et projets en France et ailleurs préparent le terrain pour une appropriation de la problématique et pour la création d'outils pour les professionnels, administrateurs et hébergeurs des sites.

Sommes-nous pour autant à la veille d'une exploitation systématique et plus large de cette information, avec diffusion des résultats ? Il est trop tôt pour le dire. La question est posée. Certaines réalisations dans le domaine du libre accès font preuve de la prise de conscience nécessaire[28]. ORI-OAI annonce des statistiques pour la version 1.6, avec un impact direct sur les projets d'archives institutionnelles des universités. Sur le plan européen, OpenAIRE[29] travaille à la mise en place d'une infrastructure du libre accès avec des statistiques d'usage. Hitchcock (2007), Salo (2008) et d'autres plaident pour la création d'un environnement de service autour des archives ouvertes.

Sur les sites des archives ouvertes en France, nous n'avons pas trouvé de services à valeur ajoutée comparables à ceux de la Public Library of Science ou de Revues.org. Cela reste, du moins en 2009, encore une utopie.

L'absence d'information sur l'accès aux sites à vocation nationale – HAL, PERSEE, Gallica pour ne citer que celles-ci – intrigue. Mais on peut supposer que la situation de 2009 évoluera vers davantage de transparence et de qualité de service.

[28] Comme notamment l'accès aux statistiques sur Revues.org.
[29] http://www.openaire.eu/index.php

5.2. Intérêt et limites de l'analyse des fichiers log

L'analyse des traces laissées par les internautes est une méthode pertinente et relativement simple à mettre en œuvre pour mesurer les pratiques sur Internet. Les études de CIBER basées sur cette « deep log files analysis » ont contribué depuis plusieurs années à mieux comprendre la réalité de la révolution numérique dans le domaine de l'information scientifique. Mais comme toutes les méthodes, celle-ci a des limites.

Borchert & Richardson (2007) pointent plusieurs faiblesses de cette méthode : des fichiers log non fiables et incomplets, des données biaisées par l'activité des robots, par "proxy caching" et par l'attribution d'adresses IP dynamiques etc. Leur conclusion est sans appel : « (Web logs) provide at best a rough estimation of repository usage » - une estimation approximative plutôt qu'une analyse fidèle.

A ceci s'ajoute une terminologie non stabilisée qui ne facilite pas l'interprétation des résultats : « Web sites now routinely measure 'hits'… with no consensus on what comprises a 'hit' » (Westell, 2006). Le fait que la plupart des concepts soit en anglais (parfois mal traduit en français) n'arrange rien. En résumé : on ne sait pas exactement ce qu'on mesure, et ces mesures ne sont pas toujours fiables.

Face au grand nombre des données, Aguillo (2009) rappelle que toutes les statistiques ne sont pas pertinentes pour l'analyse des usages. Certaines analyses nécessitent un paramétrage particulier des outils d'extraction ou bien, un lien avec d'autres sources d'information. Pour finir, l'analyse de certains éléments statistiques est sensible dans la mesure où ils touchent aux données personnelles.

Dans ce contexte assez particulier, l'approche méthodologique de notre étude de cas était prudente : appliquer la grille d'analyse développée par CIBER et utiliser un outil ayant fait ses preuves (Urchin). Le résultat est mitigé. Le choix de la grille CIBER a effectivement facilité l'analyse car elle a permis de limiter d'emblée le nombre de données et de structurer le traitement et l'interprétation des statistiques. Mais l'étude de cas a également démontré les limites de l'approche CIBER, en particulier l'incohérence des données et des concepts[30] :

1. Incohérence des données - les définitions et interprétations de CIBER changent d'une étude à l'autre. L'échantillon des données varie, certaines données n'étant pas présentes dans toutes les études ; d'autres données sont exploitées de façon différente rendant les comparaisons impossibles, comme par exemple la thématique des revues reprise des sites des éditeurs sans mise en cohérence.

2. Incohérence aussi des concepts - la distinction entre trois types de "metrics" présentée à Lille en 2009 ne correspond pas à l'approche en 2005 ou 2007 où tous les "activity metrics" (dont le nombre de consultations) faisaient partie de la catégorie "information seeking behaviour". Cela laisse penser à une conceptualisation assez faible et superficielle.

3. Incohérence des interprétations - certains éléments ne sont pas ou peu interprétés. Parfois aussi les résultats sont assez peu compréhensibles, sans explication. Certaines données paraissent "sur-interprétées", comme, par exemple, le domaine des revues interprété comme indicateur pour les disciplines scientifiques du chercheur-internaute. Comme si l'usage d'une revue classée "historique" était réservé aux seuls chercheurs en histoire.[31] D'autres données ne paraissent pas nécessairement pertinentes. Par exemple, on sait aujourd'hui que l'internaute lambda navigue "à la Google", avec un à trois mots de recherche ; pourquoi donc inclure le le nombre de termes par requête dans l'analyse des fichiers log d'un site particulier ?

[30] Nous parlons ici des publications de Nicholas et al. (2005), (2007), (2008), (2009a) et (2009b).
[31] Ce qui est faux, cf. les résultats de l'analyse par Boukacem-Zeghmouri (2010).

Les données de CIBER et d'Urchin ne se superposent pas. Certaines données CIBER ne sont pas extraites par Urchin. En contrepartie, Urchin extrait des statistiques que la grille de CIBER ne contient pas. Et comme évoqué plus haut, plusieurs mesures de la grille CIBER nécessitent un lien vers d'autres sources d'information, notamment le fichier des lecteurs (clients, utilisateurs) et les métadonnées (notices de catalogue ou base de données).

Certains problèmes ont déjà été évoqués avec les résultats (cf. 4.). Voici une synthèse.

5.3. Consultation et téléchargement

Les indicateurs d'activité ou de trafic occupent une place centrale et incontournable dans l'analyse des ressources en ligne. Cependant, l'expérience de l'étude de cas pose plusieurs questions :

Que mesure-t-on réellement ? Comment définir les objets de mesure, comment traduire les termes en français ? Il reste encore suffisamment de flou autour de la terminologie (views, visits, sessions, downloads, hits, source etc.) pour appeler de ses vœux un glossaire bi- ou multilingue.

Quel est l'indicateur central ? Le réseau DINI (projet OA Statistik) favorise l'analyse des téléchargements de documents (*downloads*) comme meilleure mesure d'impact (Henneberger, 2009). CIBER par contre considère la consultation (*page views*) comme principal indicateur de l'usage et directement lié à la taille et l'activité scientifique d'une université (Nicholas et al, 2007), avec le nombre de visites et de sessions. Ces indicateurs du trafic sur un site Web sont assez fortement corrélés ; voici le tableau des coefficients de corrélation (Spearman) pour IRIS :

Nb téléchargements	**Nb sessions**	
0,59	0,76	**Nb pages**
	0,84	**Nb téléchargements**

Tableau 32 : IRIS : Corrélation entre indicateurs d'activité (2009)

La comparaison de la distribution mensuelle de ces trois données confirme leur proximité (cf. figure 6) :

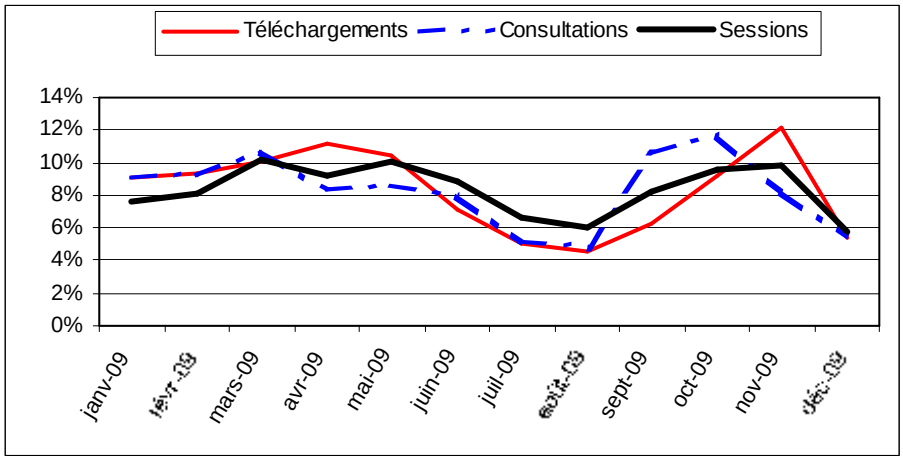

Figure 6 : IRIS : Courbes mensuelles des indicateurs d'activité (2009)

Par conséquent, tout cela mesure à peu près la même chose et il n'est peut-être pas nécessaire de tout mesurer. En moyenne, on compte entre 9 et 15 pages consultées ou visualisées par session, et entre 5 et 8 documents téléchargés. La donnée la moins pertinente ou peut-être plutôt, la plus spécifique, semble être le nombre de téléchargements. Mais on peut s'interroger sur l'effet de navigation sur cet indicateur, aussi sur son lien réel vers la consultation et la lecture du document.

Les statistiques sont-elles fiables ? Nous avons du corriger une anomalie (cf. téléchargements) sans avoir trouvé ou obtenu une explication satisfaisante. Il peut y avoir des doutes (avec Borchert & Richardson, 2007) sur la fiabilité de l'ensemble de données mais nous n'avons pas d'estimation sur le taux d'erreur. Pour la durée d'une session, CIBER explique certaines anomalies par la distribution biaisée des données et conseille l'utilisation d'une M-estimation de Hubert, une statistique moins impactée par la forme de la courbe[32] qui correspond globalement à la médiane.

Nous l'avons déjà évoqué plusieurs fois : le logiciel Urchin filtre et élimine l'activité des robots (indexation par moteurs de recherche). Le résultat est que les statistiques d'utilisation "nettoyées" sont inférieures aux chiffres obtenus jusqu'alors sans filtrage : "Usage statistics, stripped of robot usage have resulted in significantly lower figures but are now credible, compatible and consistent" (Bevan & Needham, 2009).

Comment par contre filtrer l'usage interne ? L'équipe du SCD intervient régulièrement dans le système IRIS pour déposer de nouveaux documents, ajouter ou modifier de métadonnées, éventuellement aussi pour retirer certains documents. Cette activité laisse des traces sur le serveur, et l'activité professionnelle se mêle à celle des robots et utilisateurs « ordinaires ». Urchin permet le filtrage à partir des adresses IP. Voici quelques éléments pour mesurer l'importance réelle de cette activité :

Consultation des pages	12%
Sessions	4-5%
Téléchargements	1%

Tableau 33 : IRIS : Usage interne (2009)

Ces chiffres n'ont pas été vérifiés mais paraissent logiques : moins de sessions que les utilisateurs-lecteurs externes, pratiquement pas de téléchargements mais un nombre relativement important de pages visualisées. L'indicateur de la profondeur de la visite corrobore l'image : tandis que la moyenne des pages consultées par session se situe au-dessous 10 pages pour l'utilisateur lambda, il est autour de 30 pages pour l'utilisateur professionnel qui intervient dans le système.

Quels sont les indicateurs « à problème » ? Nous n'avons pas de chiffre fiable pour le nombre des visiteurs (ou visites) répétés. Cet indicateur joue un rôle certain pour le commerce électronique : "Repeat visits to your website by your customers are the most important factor to increasing your conversion rate." (Nicholas et al., 2005). Nous n'avons pas trouvé d'équivalent dans Urchin.

[32] Cf. http://fr.wikipedia.org/wiki/M-estimateur

5.4. Recherche et navigation

Mesurer les comportements de recherche et navigation vers et sur le site est quelque peu paradoxal. Pour l'étude scientifique de l'usage des ressources en ligne, ces indicateurs sont indispensables. CIBER ne s'est pas trompé : les « *information seeking characteristics* » correspondent au plus grand nombre d'indicateurs dans leur modèle de l'analyse des fichiers log (cf. tableau 1). Pour comprendre le comportement des internautes et aussi, développer de nouveaux produits et services en adéquation de ces comportements, il faut le détail de ces mesures.

Par contre, pour avoir une idée de l'activité sur le site, il n'est pas nécessaire de mesurer le comportement virtuel avec autant de précision. C'est par ailleurs dans ce domaine-là que le logiciel Urchin paraît le moins performant. Ceci pour une simple raison : pour extraire des statistiques sur les *items* consultés, il faut des éléments descriptifs (métadonnées) pour ces *items*. Et ces éléments ne sont pas inclus dans les fichiers log mais devraient être exportés du catalogue ou de la base de données d'IRIS.

Résultat : nos données sur l'âge des documents, sur leurs disciplines etc. restent à ce jour très rudimentaires et peu satisfaisantes. Quelques analyses pourraient se faire manuellement, comme par exemple l'analyse des fichiers correspondant aux documents (*handle* vs. *bitstream*, cf. plus haut). De nouveau, c'est source d'erreur et fastidieux.

D'autres questions : Quel est le lien avec le nombre de documents déposés dans IRIS ? Quelle est la partie non utilisée ? Comment se distingue-t-elle par rapport aux documents utilisés ? Pour donner une réponse, une seule option : il faudrait établir les correspondances *handle/bitstream*-référence bibliographique pour chaque document afin de remplacer les URL par les noms (titres) des documents.

Par rapport à la navigation, nous avons déjà évoqué certains problèmes plus haut. Ajoutons ici que pour comprendre ce comportement, il faudrait créer le lien entre ces indicateurs et les caractéristiques du site en question, avec son ergonomie de son interface, son usabilité, son référencement. Ces analyses existent pour d'autres types de site Web ; il faudrait essayer de les adapter aux archives ouvertes.

5.5. Les utilisateurs

Qu'est-ce qu'un utilisateur ? D'après CIBER, tout d'abord une machine : "User identification was based on a combination of "IP" number and browser details. A user was effectively a computer; sometimes that computer represents an individual (i.e. a professor in his office), in other cases a number of people (i.e. students in the library). User background data (on occupation, organizational affiliation and geographical location) held on a registered user database was related, via an identification number, to the usage logs generated (...) by registered users." CIBER génère ces informations à partir du fichier des utilisateurs enregistrés et déclarés auprès de l'éditeur (Nicholas et al., 2005). Sans cela, l'analyse paraît compromise.

Le problème avec les utilisateurs est que l'utilisation du site IRIS ne nécessite pas d'enregistrement préalable. Seulement 22% arrivent sur IRIS via le site de Lille 1, et seulement 5% après authentification dans l'ENT (LDAP). Le reste, ce sont des « inconnus », et seule l'analyse des fichiers log avec les adresses IP et les domaines peut nous renseigner.

C'est la même situation qu'ont rencontrée Nicholas et al (2007) qui n'avaient que l'adresse IP du consortium OHIOLink. Résultat : ils ont du se contenter d'une caractérisation de l'institution identifiée ("research vs teaching university", largish vs mid-sized vs small). Mais

nous l’avons vu, l’exploitation des IP n’est pas exhaustive, ni fiable et ne permet qu’une description sommaire et partielle des disciplines et des établissements.

Nous avons donc du abandonner l’analyse de trois indicateurs : le statut professionnel, quelques informations démographiques (sexe, âge etc.), et le secteur d’activité. Mais comme indiqué plus haut, l’analyse des autres données ne produit pas non plus les résultats escomptés. Une solution est difficile à imaginer dans l’environnement du libre accès et d’archives ouvertes, par définition sans restriction d’accès.

6. 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.

6.1. Pour développer l'analyse des statistiques d'utilisation des archives ouvertes

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 de la même information ; 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 du libre accès 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 (« Archive Report 1 »).[33]

3. **Fichiers log** : De même, il faudra déterminer un nombre restreint d'éléments minimums 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 (cf. plus loin).

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 des outils et initiatives ont recours à l'anglais. Il y aura donc aussi un travail de traduction.[34]

5. **Périodicité** : Les statistiques devraient être établies sur la base d'une période mensuelle, avec un cumul annuel, et diffusées dans les 30 jours après la fin du mois.

6. **Texte intégral** : Les statistiques d'utilisation doivent différencier l'accès à un 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.

6.2. Pour développer des services à valeur ajoutée

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[35] ou du projet DINI. Cet environnement pourrait inclure :

---

[33] 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.

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

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

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).

6.3. Pour développer l'analyse des fichiers log

A partir de notre étude de cas, nous avons établi une première proposition d'indicateurs significatifs à exploiter systématiquement dans l'analyse des fichiers log. Cette liste suit les recommandations du JISC[36] et de Cybermetrics[37].

Trafic (*activity metrics, visits, downloads*)

La première catégorie d'indicateurs contient quatre données qui paraissent assez fiables et significatives pour le trafic sur un site Web tel qu'une archive ouverte :

1. Consultations (*page views*)
2. Téléchargements (*downloads*)
3. Sessions (*sessions*)
4. Profondeur de visite (*site penetration*) *

Visites (*information seeking behaviour, visits*)

Quatre indicateurs décrivent le contenu (*content*) que les internautes ont consulté :

1. Documents consultés, par exemple une liste des 20 documents les plus consultés (*name of source used*) *
2. Discipline ou thématique, c'est-à-dire le domaine des documents consultés (*subject of source*) *
3. Année de publication en trois catégories : courant, récent, archive (*age of source used*) *
4. Type des pages consultées, par exemple sommaire, résumé, texte intégral etc. (*type of material used*) *

Trois autres indicateurs permettent de mieux cerner la navigation vers et sur le site :

1. Approche de recherche : simple, avancée... (*search approach*) *
2. Mode de navigation : feuilletage, recherche sur site... (*form of navigation*) *
3. Chemin d'accès : par site de bibliothèque, via lien, Google, caché, inconnu... (*referral links*) *

[36] JISC Usage Statistics Review (2008)
[37] Aguillo (2009), Aguillo et al. (2010)

Utilisateurs (*user characteristics, visitors*)

L’intérêt de cette dernière catégorie dépend en grande partie de la qualité des données, notamment de l’accès à un fichier d'utilisateurs (données d'enregistrement). Voici le minimum, à partir des fichiers log :

1. Origine géographique, par pays (*geographical location*) *
2. Organisme employeur ou de formation (*name of organization*) *

L’information sur les utilisateurs est également limitée par le respect de l'identité humaine, des droits de l'homme, de la vie privée et des libertés.

En résumé, nous recommandons pour une analyse scientifique des fichiers log treize indicateurs dont dix statistiques cumulatives annuelles (marquées d'un *) et trois statistiques mensuelles.

7. Conclusion

Le nombre des archives ouvertes augmente continuellement et témoigne du succès d'une politique active en faveur du libre accès à l'information scientifique, malgré parfois un manque de transparence ou de cohérence sur le terrain. Leurs contenus s'amplifient également même si on peut déplorer le mélange entre les documents, les notices sans texte intégral, et les archives à caractère patrimonial.

Sur le terrain pourtant les archives ouvertes restent peu connues et semblent peu utilisées (cf. Boukacem-Zeghmouri, 2010). Dans certains cas, l'usage est absorbé par d'autres sites. La communauté des chercheurs en physique nucléaire par exemple préfère toujours Arxiv à son site miroir français devenu archive nationale, HAL. Dans les domaines des sciences de la vie et médicales, l'affaire est encore plus simple dans la mesure où un site comme PubMed Central ou UK PubMed Central n'ont pas de « concurrent » en France à ce jour.

Que dire de l'utilisation des archives ouvertes en France ? Suit-elle le développement des sites et contenus ? Ou bien doit-on s'attendre à un effet de tassement ?

Notre étude montre qu'il est aujourd'hui (encore) impossible d'avoir une idée globale sur l'utilisation réelle des archives ouvertes en France. Il y a très peu d'information publique sur l'usage fait par les chercheurs, en termes de consultation, session, navigation etc.

Il manque des outils, il manque une terminologie et des procédures, et il manque tout simplement aussi un cadre d'interprétation et d'exploitation de ces données. Cette situation peut être interprétée de plusieurs manières : l'absence de ressources humaines et technologiques, la priorité donnée au développement des sites, parfois sans doute aussi une absence volontaire de transparence dans un paysage controversé. Le résultat est le même : contrairement à la partie "commerciale" de l'offre documentaire, il est aujourd'hui impossible d'analyser l'usage des archives ouvertes, d'évaluer leur acceptation et leur utilité pour et par les chercheurs et d'avancer dans l'étude des coûts et bénéfices (retour sur investissement). Sans connaître et comprendre les usages, comment développer de nouveaux services à valeur ajoutée autour des archives ouvertes, aussi bien pour le dépôt que pour la recherche et l'exploitation ? Comparée à d'autres pays, il s'agit d'une situation anormale et certainement, transitoire.

Notre étude a établi une série de recommandations pour faciliter le suivi et l'analyse des usages, à partir des statistiques d'utilisation des archives ouvertes et notamment des fichiers log. Nous essayerons d'affiner ces recommandations dans les mois à venir en ajoutant un glossaire bilingue (anglais-français), probablement en étroite collaboration avec des équipes anglaises, allemandes et japonaises qui travaillent sur la même question : comment simplifier et normaliser l'analyse d'usage des archives ouvertes ?

8. Bibliographie

I. Aguillo (2009). `Measuring the institution's footprint in the web'. *Library Hi Tech* **27**(4):540-556.

I. F. Aguillo, et al. (2010). `Indicators for a webometric ranking of open access repositories'. *Scientometrics* **82**(3):477-486.

C. Aubry & J. Janik (2005). *Les archives ouvertes, enjeux et pratiques*. Tec Doc.

M. H. Bath (2009). `Open access repositories in computer science and information technology: an evaluation'. *IFLA Journal* **35**(3):243-257.

B. Bayer-Schur, et al. (2009). `Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers'. Tech. rep., PEER Project.

C. Bernardoni (2008). `Etude des parcours individuels de consultation à partir des logs du Conservatoire Numérique des Arts et Métiers. Quel(s) usage(s) pour le CNUM, quel(s) impact(s) pour cette bibliothèque numérique ?'. Mémoire de stage de licence professionnelle "Bibliothèques documentation et archives numériques", IUT 2 de Grenoble.

C. Bertignac & D. Gac (2009). `La voie verte : les archives ouvertes'. In *Open Access Week. 19 October 2009*. Institut Universitaire Européen de la Mer.

E. Bester & M. Dacos (2010). `Que savons-nous de l'identité, des comportements et des attentes des lecteurs de Revues.org en 2008 et 2009?'. Tech. rep., CLEO.

S. Bevan & P. Needham (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*.

B.-C. Björk, et al. (2010). `Open Access to the Scientific Journal Literature: Situation 2009'. *PLoS ONE* **5**(6):e11273+.

L. Björnshauge (2006). `The Next Generation of IRs – enabling closer cooperation & networking'. In *International Workshop on institutional repositories and enhanced and alternative metrics of publication impact. Berlin, Humboldt University. 20-21 February 2006*.

M. Borchert & J. Richardson (2007). `Usage metrics for open access repositories'. In *AusWeb07 : Sic Erat in Fatis, 30 June-4 July 2007, Novotel Pacific Bay Resort, Coffs Harbour, N.S.W.*

C. Boukacem-Zeghmouri (2010). `Pratiques de consultation des revues électroniques par les enseignants chercheurs : les STM en France'. *Documentaliste - Sciences de l'Information* **47**(2):4-13.

C. Boukacem-Zeghmouri & J. Schöpfel (2005). `Statistiques d'utilisation des ressources électroniques en ligne : le projet COUNTER'. *Bulletin des Bibliothèques de France* **50**(4):62-66.

T. Brody (2006). `Open Access Citation Index Services'. In DINI International Workshop on Institutional Repositories and Enhanced and Alternative Metrics of Publication Impact, 20-21 February 2006, Humboldt University.

C. Bruley, et al. (2007). `Résultats de l'enquête sur les projets d'archives ouvertes de la recherche dans les établissements du consortium Couperin'. Tech. rep., GTAO Couperin.

L. Carr & T. Brody (2006). `Interoperable Respository Statistics'. In *International Workshop on institutional repositories and enhanced and alternative metrics of publication impact. Berlin, Humboldt University. 20-21 February 2006*.

L. Carr, et al. (2008). `Repository Statistics: What Do We Want to Know?'. In *Third International Conference on Open Repositories 2008, 1-4 April 2008*.

M.-F. Claerebout (2003). `Grisemine, a digital library of grey university literature'. In *Fifth International Conference on Grey Literature : Grey Matters in the World of Networked Information, 4-5 December 2003*.

P. M. Davis & M. J. Fromerth (2007). `Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles?'. Scientometrics 71(2):203-215.

M. Durand-Barthez (2009). `Tendances actuelles en bibliométrie : panorama des ressources, évolution, perception'. *Documentaliste - Sciences de l'Information* **46**(4):44-49.

J. Fry, et al. (2009). `PEER Behavioural Research: Authors and Users vis-à-vis Journals and Repositories. Baseline report'. Tech. rep., LISU Loughborough University.

GFII (2009). `L'édition scientifique française en sciences sociales et humaines'. Tech. rep., Groupement Français de l'Industrie de l'Information GFII.

I. Gouat (2008). `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. 3 April 2008*.

J. Harrington & M. Betts-Gray (2009). `The Embed Project: Final Report'. Tech. rep., JISC & Cranfield University.

S. Henneberger (2009). `Standardisierte Nutzungsanalysen als alternative Impact-Messungen wissenschaftlicher Publikationen'. In *3. Open-Access-Tage, Konstanz, 7 - 8 October 2009*.

S. Hitchcock (2007). `Community EPrints: Final Report.'. Tech. rep., JISC.

S. Jana & S. Chatterjee (2004). `Quantifying Web-site visits using Web statistics: an extended cybermetrics study'. *Online Information Review* **28**(3):191-199.

J.-F. Lutz (2009). `Le mouvement pour le libre accès aux publications scientifiques'. In P. Carbone & F. Cavalier (eds.), *Les collections électroniques, une nouvelle politique documentaire*, pp. 75-85. Electre Edition du Cercle de la Librairie.

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

F. Merceur (2007). `Gestion d'une archive et d'un moissonneur, l'exemple de l'IFREMER'. In *RPIST 2007, 20 juin 2007*.

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

C. Merk & N. K. Windisch (2008). `JISC Usage Statistics Review. Final Report'. Tech. rep., JISC.

MESR (2010). `Open Access in France. A state of the Art Report - April 2010'. Tech. rep., Ministère de l'Enseignement Supérieur et de la Recherche.

D. Metje (2009). `Erste Ergebnisse der Auswertung von Nutzungsdaten'. In *3. Open-Access-Tage, Konstanz, 7 - 8 October 2009.*

S. Min, et al. (2008). `TeLearn, une archive ouverte multilingue dans le domaine des technologies pour l'apprentissage'. *AMETIST* (2).

D. Nicholas, et al. (1999). `Cracking the code: web log analysis'. *Online Information Review* **23**(5):263-269.

D. Nicholas, et al. (2005). `Scholarly journal usage: the results of deep log analysis'. *Journal of Documentation* **61**(2):248-280.

D. Nicholas, et al. (2007). `Diversity in the Information Seeking Behaviour of the Virtual Scholar: Institutional Comparisons'. *The Journal of Academic Librarianship* **33**(6):629-638.

D. Nicholas, et al. (2008). `User diversity: as demonstrated by deep log analysis'. *The Electronic Library* **26**(1):21-38.

D. Nicholas, et al. (2009a). `Online use and information seeking behaviour: institutional and subject comparisons of UK researchers'. *Journal of Information Science* **35**(6):660-676.

D. Nicholas, et al. (2009b). `Digital consumers: Virtual Scholars'. In *Ressources électroniques académiques: mesures et usages. Colloque international. Lille, 26-27 novembre 2009*.

D. Salo (2008). `Innkeeper at the Roach Motel'. *Library Trends* **57**(2):98-123.

R. Schimmer (2006). `Web Citation Index – The IR Perspective'. In DINI International Workshop on Institutional Repositories and Enhanced and Alternative Metrics of Publication Impact, 20-21 February 2006, Humboldt University.

J. Schöpfel & H. Prost (2010). `Développement et Usage des Archives Ouvertes en France. Rapport. 1e partie : Développement'. Tech. rep., Université Charles-de-Gaulle Lille 3.

P. T. Shepherd (2009). `COUNTER in context'. In *UK Serials Group Annual Conference.*

J. Sicot (2008). `Bilan de l'utilisation de HAL par la communauté scientifique de l'Ecole Centrale de Lyon'. Tech. rep., Conseil Scientifique du 10 juillet 2008.

K. Smith (2008). `Institutional Repositories and E-Journal Archiving: What Are We Learning?'. *Journal of Electronic Publishing* **11**(1).

D. Taraborelli (2005). `Institutnicod.org. Rapport sur l'impact du site sur la visibilité du laboratoire (2001-2005)'. Tech. rep., Institut Jean Nicod.

M. Westell (2006). `Institutional repositories: proposed indicators of success'. *Library Hi Tech* **24**(2):211-226.

J. Willinsky (2005). *The Access Principle: The Case for Open Access to Research and Scholarship (Digital Libraries and Electronic Publishing).* The MIT Press.

R. D. Wilson (2010). `Using clickstream data to enhance business-to-business web site performance'. *Journal of Business & Industrial Marketing* **25**(3):177-187.

Annexe A – Publications

J. Schöpfel & H. Prost (2009). `Les statistiques d'utilisation d'archives ouvertes. Etat de l'art'. In *Ressources électroniques académiques: mesures et usages. Colloque international. Lille, 26-27 novembre 2009*.

J. Schöpfel, et al. (2009). `Usage of grey literature in open archives'. In *Eleventh International Conference on Grey Literature: The Grey Mosaic: Piecing It All Together. Washington D.C., 14-15 December 2009.*

J. Schöpfel & C. Boukacem-Zeghmouri (2010). `Assessing online usage'. *Research Information* (47):25.

J. Schöpfel & H. Prost (2010a). `Développement et Usage des Archives Ouvertes en France. Rapport. 1e partie : Développement'. Tech. rep., Université Charles-de-Gaulle Lille 3.

J. Schöpfel & H. Prost (2010b). `Développement et Usage des Archives Ouvertes en France. Rapport. 2e partie : Usage'. Tech. rep., Université Charles-de-Gaulle Lille 3.

J. Schöpfel & C. Boukacem-Zeghmouri (2010). `Assessing the Return on Investments in GL for Institutional Repositories'. In D. Farace & J. Schöpfel (eds.), *Grey Literature in Library and Information Studies*. De Gruyter Saur.

H. Prost, et al. (2010 forthcoming). `Usage assessment of an institutional repository : A case study'. In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues. Prague, 6-7 December 2010*.

J. Schöpfel & C. Boukacem-Zeghmouri (2010 forthcoming). `A propos de l'analyse des usages en ligne. Notes de lecture'. *Etudes de communication* **35**.

C. Boukacem-Zeghmouri, et al. (forthcoming). `Mesures d'usage. Usage effectif et utilisabilité des sources d'information'. In E. Delamotte & G. Lallich-Boidin (eds.), *Mesure de la science.* CNRS Editions.

Annexe B – Projets, initiatives, services

IFABC

Parmi les indicateurs recommandés par l'organisation "International Federation of Audit Bureaux of Circulations"[38] figure aussi une définition d'indicateurs d'usage d'un site web, avec une terminologie normalisée pour « utilisateur », « visite » etc.

SURF-SURE

La fondation néerlandaise SURF finance un projet de normalisation, « Statistics on the Usage of Repositories » (SURE)[39], pour faciliter entre autre l'agrégation des fichiers log issus des archives ouvertes.

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[40]. 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).

Embed

Le projet Embed[41] a évalué certains outils pour l'analyse du trafic sur un site web (Google Analytics, IRStats, Minho stats pour DSpace) et a développé un outil spécifique (PERL, PHP scripts pour une base MySQL) dont les mesures sont compatibles avec COUNTER : téléchargements par collection, auteur, document (cf. Harrington & Betts-Gray, 2009).

DSpace

DSpace propose une application complémentaire (*add-on*) pour extraire des statistiques de la configuration standard[42]. Le "Stats System" de DSpace a été lancé en mars 2009, après un développement spécifique par l'université de Minho au Portugal. Parmi ses fonctionnalités, on trouve :

Personnalisation des statistiques, interfaces et accès/diffusion.
Stockage des statistiques dans une base de données.
Analyse de l'origine géographique des visites.
Identification des robots.
Statistiques en temps réel et en différé.
Agrégation des données par période.
Agrégation des données par pays, type de document etc.
Sélection des documents les plus consultés.
Interface multilingue.

[38] http://www.ifabc.org/
[39] http://www.surffoundation.nl/nl/projecten/Pages/SURE.aspx
[40] http://www.peerproject.eu/
[41] http://cclibweb-2.dmz.cranfield.ac.uk/embed/index.php/Embed_Wiki
[42] http://www.dspace.org/Add-ons/#statistics

RePEc

L'archive ouverte en sciences économiques RePEc produit des données d'utilisation d'une grande qualité[43]. 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.

NEEO

Le réseau NEEO EconomistsOnline met en oeuvre un échange de données statistiques d'usage dans un réseau d'archives institutionnelles, en conformité avec le JISC projet "Usage Statistics Review", avec la définition d'un Scholarly Works Usage Community Profile (SWUP) pour utiliser des concepts OpenURL Context Object pour les statistiques d'utilisation[44].

Les statistiques sont fournies par documents, auteurs et institutions ; elles sont modulables. Le prototype NEEO inclut aussi une sorte de "hit-parade" des documents les plus visités. Le "relevance ranking" des résultats d'une recherche est basé sur la fréquence des téléchargements (downloads).

CiteBase

Le service CiteBase reprend les références de la base ArXiv, spécialisée dans les domaines de la Physique, des Mathématiques, des Sciences cognitives et de la Biologie quantitative[45].

CiteBase établit un rapport entre le nombre de fois où un article lisible sur le Web est « ouvert » et le nombre de fois où il est cité, analyse le temps de latence situé entre la date d' « ouverture » et la date de « citation », trace une courbe marquant l'évolution d'un article à partir des relevés effectués sur le miroir britannique de la base ArXiv, suit le cycle des « avatars » d'un article et de la gerbe d'articles qu'il a pu susciter, en reliant les trois facteurs du nombre d'ouvertures, du nombre de citations et du temps intermédiaire de latence, et tient compte du nombre factice de déchargements dus à des « alertes » ou « profils » qui suscitent une ouverture quasi automatique du texte dans les premières 24 ou 48 heures (Durand-Barthez, 2009).

Open Repository

Open Repository[46] est un service proposé par BioMed Central à destination des institutions, pour leur faciliter la création d'archives institutionnelles. Ce système contient entre autre la fonctionnalité de visualiser pour chaque document les statistiques de téléchargements du résumé et du texte intégral, ainsi que des statistiques synthétiques par communauté et collection. Les auteurs peuvent recevoir les statistiques pour leurs propres dépôts par email ; l'administrateur a accès à l'ensemble des données, avec des rapports mensuels.

[43] http://logec.repec.org/
[44] http://www.economistsonline.org/home?lang=fr
[45] www.citebase.org
[46] http://www.openrepository.com

PLoS

La Public Library of Science met en ligne un service à valeur ajoutée qui s'appuie sur les statistiques d'utilisation au niveau des articles (PLoS article-level metrics[47]). Ce service est conçu pour les lecteurs, comme un indicateur de la qualité ou de l'impact de l'article. Il est proposé avec d'autres éléments : les citations dans la littérature, social bookmarking, blogs et les notes, commentaires et "star ratings" des autres lecteurs (tags). Pour chaque article, le lecteur a accès à une suite de 20 données descriptives concernant l'article individuel, à partir d'une touche de fonction "metrics". L'ensemble des données peut être téléchargé sous forme de tableau. A titre comparatif, il y a également des tableaux synthétiques avec les statistiques pour une catégorie d'articles, classés par âge, discipline et revue[48].

Ce service est évolutif et sera enrichi par les données en provenance d'autres sources. Voici la liste des données actuelles :

Données descriptives :

1. DOI - The Digital Object Identifier (DOI) is a unique identifier for academic articles. Publication Date - the official publication date of the article.
2. Publication Year - the official year of publication of the article
3. Journal - the name of the PLoS Journal
4. Research Article' or 'non-Research Article'?
5. Article Title - this is an indication of the title of each article.

Données d'usage et d'impact :

1. Citations - CrossRef - this is a count of how many citations are recorded to this article by CrossRef
2. Citations - PubMed Central - this is a count of how many citations are recorded to this article by PubMed Central
3. Citations - Scopus - this is a count of how many citations are recorded to this article by Scopus.
4. Total HTML Page Views - this is a total number for the HTML page views to each article
5. Total PDF Downloads - this is a total number for the PDF downloads for each article
6. Total XML Downloads - this is a total number for the XML downloads for each article
7. Combined Usage (HTML + PDF + XML) - this is the sum of the previous three numbers.
8. Blog Coverage - Postgenomic - this is a count of how many blog postings refer to this article, as indexed by the Postgenomic service
9. Blog Coverage - Nature Blogs - this is a count of how many blog postings refer to this article, as indexed by the Nature Blogs service
10. Blog Coverage - Bloglines - this is a count of how many blog postings refer to this article, as indexed by the Bloglines service
11. Number of Trackbacks - this is an indication of the number of trackbacks that have been made to this article by external sites
12. Social Bookmarking - CiteULike - this is a count of how many bookmarks have been made to this article by users of the CiteULike social bookmarking service
13. Social Bookmarking - Connotea - this is a count of how many bookmarks have been made to this article by users of the Connotea social bookmarking service
14. Number of Ratings - this is a count of how many times an article has been 'rated' on our site
15. Average Rating - ratings can be made in several categories (and in addition, older rating functionality on some titles only allowed a single rating to be made). This therefore indicates the average rating that the article has received
16. Number of Note threads - users can leave a 'Note' on a specific piece of text in any article. Once a Note has been left, it forms a discussion thread to which other users can reply. This field provides a count of how many Note threads have been started on the article.
17. Number of replies to Notes - once a Note thread has been started, users can reply in that thread. This field provides a count of how many replies to Note threads have been made, in total.

[47] http://www.plos.org/cms/node/485

[48] Cf. http://www.plos.org/about/faq.html#metrics et http://article-level-metrics.plos.org/

18. Number of Comment threads - users can leave a 'Comment' on the entire article. Once a Comment has been left, it forms a discussion thread to which other users can reply. This field provides a count of how many Comment threads have been started on the article.
19. Number of replies to Comments - once a Comment thread has been started, users can reply in that thread. This field provides a count of how many replies to Comment threads have been made, in total
20. Number of 'Star Ratings' that also include a text comment - when users leave a Rating on an article they also have the opportunity to leave a text comment. This text comment is not noted in the 'Comments' or 'Notes' functionality on our site. Therefore to get a complete picture of all comment activity you should include the comments that are left as part of a Rating. This field indicates how many comments have been left in this way.

Rian.ie – Pathway to Irish Research[49]

Mars 2010, l'enseignement supérieur irlandais a mis en place un portail d'accès à différentes archives institutionnelles d'Irlande. Ce site contient plusieurs pages de statistiques, aussi bien pour les dépôts (contenu) que pour l'utilisation (consultations, visites), aussi bien au niveau du site que de chaque document. L'accent est mis sur l'origine géographique des internautes et sur le nombre de visualisations. Ces éléments s'adressent aussi bien aux professionnels (administrateurs des sites) qu'aux lecteurs-auteurs, avec des données sur l'impact du document.

CLEO/Revues.org

Le CLEO a mené une étude sur l'usage des revues en libre accès de la plate-forme Revues.org qui réuni analyse des fichiers log (outil Awstats) et enquête qualitative en ligne (Bester & Dacos, 2010).

D'une manière exemplaire, le CLEO met à disposition en ligne les données d'utilisation de la plate-forme, en détail :

1. données mensuelles
2. par titre (y compris les blogs scientifiques)
3. nombre de visiteurs uniques
4. nombre de visites
5. nombre de pages visitées
6. nombre d'accès ("hits")
7. taille des fichiers téléchargés
8. total mensuel

Pour chaque titre (revue ou blog), le CLEO propose l'ensemble des données Awstats, c'est-à-dire :

Résumé
Quand:
Historique mensuel
Jours du mois
Jours de la semaine
Heures
Qui:
Pays
... Liste complète
Hôtes
... Liste complète
... Dernière visite
... Adresses IP non résolues
Visiteurs Robots/Spiders
... Liste complète

[49] http://www.rian.ie/en

... Dernière visite
Navigation:
Durée des visites
Types de fichiers
Pages vues
... Liste complète
... Entrée
... Sortie
Systèmes exploitation
... Versions
... Inconnu
Navigateurs
... Versions
... Inconnu
Origine/Referer:
Origine de la connexion
... Moteurs de recherche
... Sites référenceurs
Recherche
... Phrases clés
... Mots clés
Autres:
Divers
Codes Status HTTP
... Pages non trouvées

C'est unique en France à ce jour.

Annexe C – A propos du logiciel *Urchin*

Urchin fut une petite société de logiciel spécialisée dans l'analyse du trafic sur le Web avant d'être rachetée par Google en 2005. Quelques mois plus tard, Google lança le logiciel d'*Urchin* sous le nom de Google Analytics, avec une diffusion gratuite. Le succès était immédiat - plus de 250 000 inscriptions en deux jours.

Aujourd'hui, il en existe trois versions :

Google Analytics, le système de base qui peut être téléchargé gratuitement en quelques instants.

AWStats (Advanced Web Statistics), une version augmentée.

Urchin 5, le logiciel commercial avec davantage de fonctionnalités (cf. Tyler & Ledford, 2006).

Nous avons trouvé trois études qui appliquent la méthodologie de l'analyse des fichiers log en évoquant le logiciel *Urchin* (Jana & Chatterjee, 2004 ; Aguillo, 2009 ; Wilson, 2010).

Parmi les logiciels du domaine récent appelé "cybermetrics" ou webométrie, *Urchin* est considéré comme un logiciel puissant pour le secteur commercial ("business"), facile à installer et utiliser, très rapide (Wilson, 2010), avec des rapports très détaillés.

Mesurer l'empreinte d'une institution sur le Web recouvre trois approches méthodologiques, l'analyse de son activité sur Internet, l'évaluation de son impact et l'étude de son utilisation, c'est-à-dire du trafic sur son site (Aguillo, 2009). L'analyse des fichiers log contribue à ce dernier objectif, et Aguillo (2009) souligne l'intérêt et aussi le succès du logiciel d'*Urchin* - parce qu'il est puissant, diffusé et maintenu par Google. Mais il ajoute, par rapport à la version Analytics : " (...) the standard configuration is not very complete, it is difficult to customize and not very well designed for downloading analysis".

Annexe D – Glossaire

Terme	Synonyme(s)	Traduction	Définition
accès	*hit*	access	Utilisé par les compteurs de pages web et désigne une requête à un serveur HTTP demandant un fichier (image, HTML, javascript, feuille de style CSS, etc). A la dfférence avec le "hit" qui correspond à une requête pour n'importe quel fichier d'un server Web. which refers to a request for any file from a web server. There may therefore be many hits per 'page' view since an html 'file' can be made up of multiple pages.
adresse IP		IP address	Une adresse IP (avec IP pour *Internet Protocol*) est un numéro qui est attribué à chaque appareil électronique lorsqu'il participe à un réseau informatique utilisant l'Internet Protocol.
année de publication		age of source	Date ou année de publication ; parfois aussi date ou année de dépôt sur le site.
approche de recherche		search approach	Comportement sur site. Chemins de navigation les plus empruntés. Une « combination of items » en termes de stratégie de recherche. Approche de recherche : simple, avancée...
archive institutionnelle		institutional repository	Relève d'une institution et a pour objectif de contenir, valoriser et conserver l'ensemble de la production scientifique de celle-ci.
archive ouverte		open archive	Désigne un réservoir où sont déposées des données issues de la recherche scientifique et de l'enseignement et dont l'accès se veut ouvert c'est-à-dire sans barrière
authentification		authentication	Procédure qui consiste, pour un système informatique, à vérifier l'identité d'une entité (personne, ordinateur...), afin d'autoriser l'accès de cette entité à des ressources (systèmes, réseaux, applications...).
bibliothèque numérique		digital library	Un ensemble d'écrits numérisés et accessibles à distance (en particulier *via* Internet).
bitstream		bitstream	Une série de bits. Document.
clé de recherche		search key	Liste des phrases/mots clés utilisés, avec l'affichage des entrées en ordre, selon le nombre de visiteurs correspondant à chacune d'elles.
comportement sur site		on site seeking behaviour	Comportement sur site. Chemins de navigation, en termes de stratégie de recherche.
consultation	*vs. hit*	page view, use	Requête pour télécharger un fichier HTML (page) d'un site Web. On the World Wide Web a 'page' request would result from a web surfer clicking on a link on another 'page' pointing to the 'page' in question. A la dfférence avec le "hit" qui correspond à une requête pour n'importe quel fichier d'un server Web. which refers to a request for any file from a web server. There may therefore be many hits per 'page' view since an html 'file' can be made up of multiple pages.
document patrimonial		heritage item	Document avec valeur historique et culturelle.
ergonomie		ergonomics	On nomme ergonomie « l'étude scientifique de la relation entre l'homme et ses moyens, méthodes et milieux de travail » et l'application de ces connaissances à la conception de systèmes « qui puissent être utilisés avec le maximum de confort, de sécurité et d'efficacité par le plus grand nombre »
feuille de style		style sheet	C'est un document permettant de mettre en forme un autre document, comme par exemple un document rédigé dans un langage de balisage, comme SGML, HTML ou XML. Pour ce faire les feuilles de style sont rédigées dans un langage spécifique.
feuilletage		browsing	A web browser or Internet browser is a software application for retrieving, presenting, and traversing information resources on the World Wide Web. An *information resource* is identified by a Uniform Resource Identifier (URI) and may be a web page, image, video, or other piece of content. Hyperlinks present in resources enable users to easily navigate their browsers to related resources.

fichier log	*journal*	log file	le concept d'historique des événements ou de logging désigne l'enregistrement séquentiel dans un fichier ou une base de données de tous les événements affectant un processus particulier (application, activité d'un réseau informatique…). Le journal (en anglais *log file* ou plus simplement *log*), désigne alors le fichier contenant ces enregistrements. Généralement datés et classés par ordre chronologique, ces derniers permettent d'analyser pas à pas l'activité interne du processus et ses interactions avec son environnement.
handle		handle	Valeur qui permet l'accès en lecture et/ou écriture à une donnée située soit en mémoire principale soit ailleurs. Une référence n'est pas la donnée elle-même mais seulement une information de localisation. La plupart des langages de programmation permettent l'utilisation de références, que ce soit de façon explicite ou implicite.
hit	*accès*	hit	L'accès à un fichier ou à une page Web.
item	*document*	item	Un objet quelconque, généralement considéré sous l'angle de sa dimension marchande. Ici : document
libre accès		open access	La mise à disposition gratuite d'un document sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral (...) ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet.
littérature grise		grey literature	Ce qui est produit par toutes les instances du gouvernement, de l'enseignement et la recherche publique, du commerce et de l'industrie, sous un format papier ou numérique, et qui n'est pas contrôlé par l'édition commerciale.
métadonnées		metadata	Ensemble de données structurées décrivant des ressources physiques ou numériques.
mode de navigation	*comportement sur site*	form of navigation	Le comportement sur site - recherche, feuilletage à partir des listes de sujets, auteurs ou documents etc.
navigateur		navigator	Un navigateur Web est un logiciel conçu pour consulter le World Wide Web. Techniquement, c'est au minimum un client HTTP.
ORI-OAI			Logiciel développé par les universités françaises pour leurs archives institutionnelles et le moissonnage OAI.
page		page	Une ressource du World Wide Web conçue pour être consultée par des visiteurs à l'aide d'un navigateur Web. Elle a une adresse Web. Techniquement, une page Web est souvent constituée d'un document en Hypertext Markup Language (HTML) (ou XHTML) et d'images. Cependant, tout type de ressources ou d'assemblage de ressources, textuelles, visuelles, sonores, logicielles, peuvent constituer une page Web.
profondeur de visite		site penetration	Nombre moyen d'accès aux pages, par visite. Ou plus simplement, le nombre de pages consultées lors d'une session.
requête	*recherche*	request, search, query	Demande de traitement. Dans le contexte des protocoles client-serveur, une requête est le message initial envoyé par le client vers le serveur. Le message du serveur étant la réponse.
robot	*crawler, spider, bot, robogiciel*	robot	Agent logiciel automatique ou semi-automatique qui interagit avec des serveurs informatiques. Un bot se connecte et interagit avec le serveur comme un programme client utilisé par un humain.
session		session	Période délimitée pendant laquelle un appareil informatique est en communication et réalise des opérations au service d'un client - un usager, un logiciel ou un autre appareil.
source	*document*	source	Documents déposés dans l'archive.
stratégie de recherche	*approche de recherche*	search style	Comportement sur site. Chemins de navigation les plus empruntés. Une « combination of items » en termes de stratégie de recherche. Approche de recherche : simple, avancée...
succès		success	Requête traitée avec succès (création d'un document etc.).
taille du document		size of source	Taille du fichier qui correspond au document.

taux de rebond		bounce rate	Le pourcentage d'internautes qui sont entrés sur une page Web et qui ont quitté le site après. Ils n'ont vu qu'une seule page.
téléchargement		full-text download	L'opération de transmission d'informations — programmes, données, images, sons, vidéos — d'un ordinateur à un autre via un canal de transmission, en général internet ou intranet.
usabilité		usability	Définie par la norme ISO 9241 comme « le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié ».
usager	*visiteur*	user	Internaute. Machine.
visite		visit	Série de pages web consultées de façon consécutive durant un laps de temps défini. On parle aussi de sessions.
visite répétée		repeated visit	Plusieurs sessions, par le même utilisateur.
visiteur		visitor	Internaute qui consulte une série de pages web consécutives (détectées par les tags) et effectue une série de requêtes Web (enregistrées dans les logs).
visualisation d'une page	*visite, vue*	page view	Désigne le nombre de fois où une page web est affichée ('rendue') dans un navigateur web. On parle aussi d'impressions.
voie de navigation		referral link	Chemin d'accès vers le site.

Sources : Wikipédia France, site d'actualité Libre Accès de l'INIST-CNRS, Urchin, H. Prost & J. Schöpfel