Mission

Indicateurs d’usages des ressources électroniques

Rapport final

Sabine BARRAL

Chargée de mission

Mars 2007

Je remercie tous ceux qui m’ont aidée à réaliser cette étude. Il est impossible de les citer tous mais je voudrais remercier plus particulièrement le Président de l’Université de Lyon 1 qui m’a accueillie dans son université et le Directeur du Service commun de la documentation (SCD) qui m’a ouvert les portes de la bibliothèque et a montré tout au long de ces mois beaucoup d’intérêt pour ce travail tout en m’accompagnant de ses conseils judicieux.

Table des matières

Introduction

La mission « Indicateurs d'usages des ressources électroniques » (Voir lettre en annexe 1) m'a été confiée par la Direction générale de l'enseignement supérieur (DGES[1]) et plus particulièrement la Sous-direction des bibliothèques et de l'information scientifique (SDBIS). Elle a commencé le 1er janvier 2006. Cette mission étant arrivée à son terme, ce rapport tente de faire la synthèse du travail accompli.

▪ Objectifs

L'étude proposée, intitulée « Indicateurs d'usages des ressources électroniques » s'insère dans une problématique d'actualité : le thème rejoint, en effet, une préoccupation nationale, conséquence de la mise en place de la loi organique relative aux lois des finances (LOLF) et du besoin d'un système d'évaluation qui en découle, système fondé sur des indicateurs précis. L'Administration souhaite bénéficier d'indicateurs de performance pour

[1] Une liste des sigles est présentée en annexe 2

évaluer, dans le cas qui nous préoccupe, les ressources électroniques utilisées dans les bibliothèques. Cette étude rejoint aussi les recommandations de la Cour des comptes dans son rapport public de 2005 dont l'une consiste à « mieux évaluer la performance des BU à partir d'indicateurs d'efficacité et d'efficience ».

Cette étude rejoint également un besoin exprimé par les bibliothèques des établissements d'enseignement supérieur, besoin qui s'est accru au fil des mois. En effet, la croissance exponentielle des dépenses en ressources électroniques[2] (14 fois plus de dépenses en 2001 qu'en 2005), le mode de dépenses (souvent des marchés sur trois ans)… posent de nombreuses questions aux établissements universitaires. Après une période de croissance importante en achats et abonnements à des ressources électroniques (20% des acquisitions dans les BU, BIU… en 2005 contre 1,7% en 2001), plus particulièrement dans le domaine STM mais aussi en droit/économie et SHS, ce qui a induit une extension très forte de l'usage de ce nouveau support, plusieurs Services communs de la documentation (SCD) ont commencé, leurs ressources stagnant, à réduire leurs dépenses, compte tenu de l'augmentation continue des coûts de ce support. Si cette tendance se confirme, les SCD devront, plus qu'avant peut-être, disposer des outils nécessaires pour, d'une part, évaluer l'intérêt des ressources acquises et, d'autre part, justifier leurs besoins financiers auprès de leurs établissements et de leur tutelle. Pour ce faire, il est essentiel de définir des statistiques ajustées et de construire les indicateurs de performance afférents, permettant d'évaluer finement l'usage des ressources électroniques par rapport aux

[2] Informations provenant des rapports ESGBU et ERE de la SDBIS

ressources papier en termes de consultation et de coût, ou encore d'évaluer l'efficacité et l'efficience de leur usage. C'est ce besoin qui a été, par exemple, mentionné par T. Plum lors de la conférence *« International developments in library assessment and opportunities for Greek libraries technological education institution*, Thessaloniki, 2005 », T. Plum citant J.C. Bertot et D.M. Davis[3]. Pouvoir tenir à jour des tableaux de bord sur l'usage des ressources électroniques devient indispensable aux directeurs des SCD pour les aider à participer de façon argumentée à la construction de la politique documentaire de leur établissement et à définir la meilleure stratégie pour constituer les collections électroniques utiles pour leur public.

C'est également une nécessité sur le plan national, comme le soulignait également, en janvier 2006, Monsieur Jolly, alors directeur de la SDBIS, lors de la réunion des directeurs de bibliothèques, précisant que, malgré l'importance des statistiques, les résultats remontés à ce jour par les établissements restaient très incomplets (les taux de réponse concernant les ressources électroniques dans le formulaire correspondant de l'ESGBU variaient en 2003 entre 10% et 57% en moyenne selon les indicateurs[4]).

Et pourtant, l'arrivée des ressources électroniques devrait faciliter la collecte des données statistiques par rapport à ce qui se passait durant l'ère « papier », puisque la trace des consultations est plus facile à suivre. Mais aujourd'hui, les SCD semblent manquer de culture statistique ou tout au moins de moyens pour collecter des données statistiques, d'où l'intérêt d'une étude qui aborde également les dispositifs techniques permettant d'effectuer

[3] "Why should libraries collect information about the usage of their networked electronic resources? As Bertot and Davis [(Planning...)] point out, there are at least two reasons : 1. To develop access to critical data that can help libraries make decisions regarding services and resources ; 2. To develop data-rich evidence for the patron communities that the library serves attesting to the value of the library-enabled networked services and resources." (In Evaluating… p. 1)

[4] Information provenant de la SDBIS

des mesures statistiques sur l'usage des ressources électroniques et de les exploiter.

Par ailleurs, l'augmentation annuelle importante des coûts des ressources électroniques mentionnée ci-dessus, en particulier pour les abonnements aux revues (souvent 6 à 7%, parfois au-delà de 10%, en tout cas plus que l'inflation en France) rend les négociations avec les fournisseurs de plus en plus serrées. Là encore, les établissements ont besoin de statistiques pour construire leurs propres modèles économiques sur des critères qui ne sont pas toujours ceux que proposent les fournisseurs à partir de leurs propres statistiques ; cela faciliterait, comme le souligne J.C. Bertot, les négociations.

La première partie de cette étude s'attache donc à proposer quelques données statistiques et indicateurs pouvant fournir aux différents gestionnaires concernés les éléments dont ils ont besoin pour mesurer la performance de l'usage des ressources électroniques dans leurs établissements.

Un deuxième objectif assigné à la mission, complément logique du premier, concerne l'étude « généraliste des dispositifs techniques » permettant [aux établissements] de « collecter et [d'] assurer l'exploitation des données destinées à construire les indicateurs »[5] attendus. Cet objectif s'est avéré plus difficile à atteindre, d'une part parce qu'il requiert des compétences techniques, d'informaticien en particulier, non prévues dans la lettre de mission ; d'autre part parce que l'expression « collecter les données » peut s'interpréter de deux façons :

[5] Selon les termes de la lettre de mission

- collecter par extraction des données brutes et mettre en place des compteurs pour effectuer les mesures nécessaires à l'obtention des données statistiques attendues, puis les exploiter ;
- collecter des données statistiques externes et trouver les dispositifs nécessaires à l'exploitation de ces données pour construire les indicateurs choisis.

Si les deux interprétations se rejoignent, la première demande un investissement technique plus lourd car elle nécessite la mise en place de plates-formes de test ; cette solution a cependant été privilégiée dans l'étude car elle donne, au moins en son principe, une indépendance aux établissements par rapport à leurs fournisseurs ; de plus, parce qu'elle apparaissait techniquement plus difficile, il était intéressant de l'analyser en profondeur.

▪ Méthodologie

Sur le plan méthodologique, ce travail a consisté, en ce qui concerne le premier objectif, en une étude bibliographique complétée par des rencontres et des mises au point lors des comités de suivi (Voir ci-dessous). Le deuxième objectif a non seulement fait l'objet d'une étude bibliographique et de rencontres, y compris avec des fournisseurs de ressources ou de logiciels et des informaticiens, mais aussi de tests destinés à éprouver la faisabilité de mesures locales de l'usage des ressources électroniques. Ces tests ont eu lieu dans trois sites (Université de technologie de Compiègne, SCD de l'Université Louis Pasteur de Strasbourg, SICD1 de l'Université Joseph Fourier et de l'Institut national polytechnique de Grenoble associé à l'Institut

d'informatique et mathématiques appliquées de Grenoble), puis en grandeur nature à l'Université Claude Bernard Lyon 1 qui souhaitait bénéficier d'ores et déjà d'un système de mesure susceptible de s'insérer dans un projet plus large de tableau de bord. Enfin, une enquête « légère » a été lancée en août 2006 auprès de 102 bibliothèques des établissements d'enseignement supérieur (à la demande du directeur de la SDBIS, l'étude n'a pas concerné les établissements de recherche) afin de connaître la disponibilité des statistiques d'usage des ressources électroniques dans les SCD et les outils techniques éventuellement déjà implantés (Voir annexe 3).

Concrètement, ce travail a été accompagné par un Comité de suivi mis en place par la SDBIS, dirigé par le Sous-directeur des bibliothèques et composé de directeurs et autres représentants des SCD et de six membres de la Sous-direction (Voir liste des membres en annexe 4). Ce comité s'est réuni quatre fois, la participation de ses membres a été active. Ces réunions ont chaque fois fait l'objet d'un envoi préalable de rapports d'étape et autres documents de travail.

A été également mis en place un comité technique chargé d'étudier les aspects plus techniques : analyser la faisabilité des tests, organiser ces tests dans les établissements, voir l'impact potentiel sur les Centres de ressources informatiques (CRI), trouver les outils nécessaires pour ces mesures… ; il ne s'est réuni qu'une fois. Une mutualisation des tests a été lancée (Voir ci-dessus) après que le comité a validé l'intérêt de travailler sur des statistiques locales.

Les pages qui suivent reprennent les objectifs définis dans cette introduction sous forme de deux chapitres : indicateurs de performance de l'usage des ressources électroniques ; dispositifs techniques pour collecter et exploiter les données statistiques. Un troisième chapitre tentera de proposer des éléments de synthèse.

Chapitre 1. Indicateurs de performance

▪ Etude bibliographique

Dans leur livre « *E-Metrics for library and information professionals* », A. White et E.D. Kamal soulignent que la méthode de capture des données d'usage détermine la nature des indicateurs pouvant être calculés. Il est effectivement difficile de séparer, dans la littérature, ce qui traite des données statistiques et des indicateurs utiles pour mesurer la performance de l'usage des ressources électroniques et des aspects techniques afférents. Mais pour la clarté du rapport, il a été décidé de séparer l'étude bibliographique en deux et de ne traiter ici que la question du choix des statistiques et indicateurs. Les éléments bibliographiques relatifs aux aspects techniques seront présentés dans le deuxième chapitre de ce rapport.

Pour définir ces données, il faut d'abord se demander, comme le souligne la littérature, dans quel but on souhaite les récolter, quels sont les enjeux et quels en sont les utilisateurs potentiels.

Ce travail a donc débuté par une étude bibliographique. De nombreux articles (Voir sélection en annexe 5) ont été écrits sur l'utilisation des ressources électroniques et en particulier sur les statistiques et indicateurs nécessaires pour évaluer ledit usage ou sur les moyens pour y parvenir : C. Tenopir[6] a répertorié 200 articles de 1995 à 2003 (H.R. Jamali). Une quinzaine d'articles rédigés depuis 2004 ont encore pu être sélectionnés, au cours de la mise à jour bibliographique très sélective qui a précédé la préparation de ce rapport.
Les travaux de l'ARL (Association for research libraries) sont particulièrement nombreux. On peut noter également les travaux des organismes ou projets tels Liber (Ligue des bibliothèques européennes de recherche), ICOLC (International coalition of library consortia), EQUINOX[7] (programme de la Commission européenne), BIX (Bibliotheksindex), COUNTER (Counting Online Usage of NeTworked Electronic Resources / Comptage de l'utilisation en ligne des ressources électroniques en réseau). S'il y a eu des études en France, elles semblent ne pas avoir fait l'objet de nombreuses publications.

L'ARL a beaucoup travaillé sur les statistiques des bibliothèques afin de définir un protocole commun nécessaire pour comparer les bibliothèques, préoccupation qui se trouve également au cœur de cette mission. Après les statistiques sur les objets « classiques », l'ARL s'est penchée sur les statistiques pour les ressources électroniques[8]. Les travaux de l'ARL ont

[6] C. Tenopir est professeur en sciences de l'information à l'Université du Tennessee
[7] Voir glossaire en annexe 6
[8] « In an environment with increasing emphasis on digital resources, what are the metrics that are appropriate for describing and characterizing collections, and what collection trends are particularly important to identify and track over time ? ; - Increasing demand for libraries to demonstrate outcomes/impacts in areas important to the institution. -Increasing pressure to maximize use of resources – benchmark best practices to save or reallocate resources». (M. Kyrillidou. – Le New Measures…, vue 11)

conduit aux projets E-metrics[9], qui cherche à faire la synthèse de tous les standards, MINES (Measuring the impact of networked electronic services) qui s'appuie sur des enquêtes et non des comptages... La situation de l'ARL est intéressante car elle rejoint les attentes françaises.

Les normes concernant les « *Statistiques internationales des bibliothèques* » (ISO 2789) et les « *Indicateurs de performance des bibliothèques* » (ISO 11620) ont été analysées soigneusement. Ces deux normes sont indispensables aux gestionnaires des bibliothèques. Il est intéressant de souligner que la première édition de la norme ISO 11620 (1998) ne comportait pas de chapitre sur les indicateurs de performance pour les services et ressources de bibliothèques électroniques. Un rapport technique ISO/TR 20983 a été publié sur ce sujet en 2003 et la nouvelle édition de la norme ISO 11620 (dans sa version soumise à enquête probatoire et qui a fait l'objet très récemment d'un vote positif) « intègre … les indicateurs de performance pour les services et les ressources de bibliothèques électroniques et traditionnelles »[10]. Les définitions présentées dans ces normes sont un élément essentiel car leur respect conditionne la possibilité d'interpréter de manière univoque les données mesurées. On trouve également dans la norme ISO 11620 une présentation détaillée des modes de calcul relatifs aux indicateurs présentés.

Dans la même ligne, l'initiative COUNTER est un projet intéressant qui a émergé à partir d'un certain nombre d'initiatives existantes et a vu le jour en

[9] Voir glossaire en annexe 6
[10] Voir l'avant-propos

mars 2002 sous l'égide d'un groupe d'éditeurs, d'acteurs intermédiaires et de bibliothécaires ou associations de bibliothécaires. En 2004, l'initiative COUNTER est devenue une entreprise non commerciale « Counter online metrics » dirigée par un Conseil d'administration, un Comité de direction et un Comité consultatif. Le président actuel est un éditeur et Peter T. Shepherd, directeur de projet, en organise le développement.

COUNTER publie un « Code de bonnes pratiques » permettant de mesurer l'utilisation des produits et services en ligne d'une manière crédible, cohérente et compatible (les 3 C de Peter T. Shepherd, directeur du projet) ; actuellement il existe un code pour les bases de données et les revues en ligne et un pour les livres et ouvrages de référence.

Ce Code de bonnes pratiques spécifie :

- les éléments de données devant être mesurés ;
- les définitions de ces éléments (certaines sont reprises dans les normes ISO 2789 et ISO 11620) ;
- le contenu, le format, la fréquence et le mode de transmission des rapports d'utilisation…

Il est consultable à l'adresse http://www.projectcounter.org/code_practice.html et sur le site de l'Institut de l'information scientifique et technique (INIST) à l'adresse http://www.inist.fr/article164.html pour la traduction française. Pour être déclarés conformes à COUNTER ou « compatibles COUNTER », les fournisseurs doivent en faire la demande auprès du directeur de projet, ils seront ensuite « audités » : les premiers audits doivent avoir lieu avant fin juin 2007, puis les fournisseurs seront « audités » une fois par an. Joint par mél à l'automne 2006, P. Shepherd indiquait qu'aucun éditeur n'avait encore été

« audité ». Une liste des fournisseurs déclarés conformes est tenue à jour sur le site de COUNTER. A ce jour, aucun éditeur français ni même francophone n'a été déclaré conforme, même si quelques-uns fournissent des rapports construits selon les recommandations COUNTER. Par contre, la liste augmente régulièrement : 26 en mai, 48 en septembre, 55 en décembre. Mais tous les fournisseurs accepteront-ils de suivre le Code COUNTER (C. Anderson) ?

La version 1 du code pour les bases de données et les revues en ligne est parue en décembre 2002, la version 2 en avril 2005 avec effet au 1er janvier 2006. Elle recommande l'élaboration de 8 rapports dont 3[11] ne sont pas obligatoires (Voir liste en annexe 7). La première version pour les livres et les ouvrages de référence est parue en mars 2006. Elle recommande l'élaboration de 6 rapports (Voir liste en annexe 7). Il est à noter que, parmi tous ces rapports, aucun n'est prévu concernant les usagers (par adresse IP par exemple). Cependant, certains éditeurs compatibles COUNTER fournissent, en plus des rapports officiels, ce type d'informations.

Le Code COUNTER donne également des directives concernant l'élaboration des rapports et donc les types de mesures. Ce point sera repris dans le chapitre suivant.

Le National Information Standards Organization (NISO), « l'Afnor américain », est associé au projet depuis le début. Les responsables du Code COUNTER souhaitent que celui-ci soit un cadre pour faciliter l'élaboration et l'échange de statistiques au niveau international. Mais si l'Association internationale de normalisation (ISO) se réfère largement aux définitions de COUNTER, seule

[11] Le rapport JR1a a été ajouté très récemment, sans qu'il y ait une nouvelle édition

l'Association nationale apporte officiellement son soutien. Il est à noter que l'ISO a décidé récemment de créer une fonction d'officier de liaison auprès de COUNTER. L'objectif de cette fonction est de veiller entre autres à une compatibilité des définitions[12].

Cette étude bibliographique s'est fortement appuyée sur la synthèse très riche *« Planning and evaluating library networked services and resources »* éditée par J.C. Bertot et D.M. Davis, J.C. Bertot étant membre du TC46/SC8 (comité technique de l'ISO) et responsable de la révision de la norme ISO 11620. Cette compilation donne, entre autres, des tableaux comparatifs sur les données statistiques et indicateurs sélectionnés par divers organismes ou autres initiatives (ISO, NISO, COUNTER, ICOLC…) et comporte les définitions afférentes : ces tableaux ont servi de référence et de modèle dans la suite de l'étude. On peut noter que les différents travaux de l'ARL mais aussi de BIX, le projet EQUINOX, n'ont pas toujours mis l'accent sur les mêmes données (coûts, satisfaction des usagers, utilisation des ressources…).

Si les premières études ont privilégié les mesures quantitatives (différentes expériences ont eu lieu entre 1998 et 2002), apparaissent ensuite également des mesures qualitatives (MINES) ; même si celles-ci sont sans doute moins fiables, elles rejoignent une préoccupation en « émergence » dans les établissements publics français, à savoir la qualité (norme ISO 9000…), la

[12] "As to COUNTER, we wanted to have a liaison because we want as much as possible to have the same definitions as COUNTER, and also because we want COUNTER to look at what we do (R. Poll). "

satisfaction des usagers… A ce jour, l'ARL semble porter ses efforts sur les projets LIBQUAL et DIGIQUAL[13].

Pour compléter cette première approche, un certain nombre de contacts (rencontres, conversations téléphoniques, messages électroniques) ont été pris (Voir liste en annexe 8). Beaucoup ont concerné les questions techniques et les conclusions seront reprises dans la partie consacrée à cet aspect.

- **Choix des items pertinents**

La lettre de mission demandait l'élaboration d' « items pertinents » concernant données statistiques et indicateurs.

Grâce aux éléments repérés lors de l'étude bibliographique, à l'étude des normes et standards et autres recommandations, grâce également aux données déjà collectées dans différentes enquêtes lancées depuis plusieurs années par la SDBIS, un état de l'art a pu être constitué. A partir de ces informations ainsi que des comptes rendus d'expériences trouvés dans la littérature, un premier tableau de statistiques et indicateurs paraissant intéressants pour évaluer l'utilisation des ressources électroniques dans les universités françaises a été dressé et soumis aux membres du Comité de suivi. Ce tableau définissait également un premier jeu d'objets sur lesquels mesurer les usages desdites ressources : bases de données, périodiques électroniques, site Web de la bibliothèque…

Ce tableau a été longuement discuté dès la première réunion du Comité. Les

[13] Voir glossaire en annexe 6

échanges ont porté sur les indicateurs tels qu'ils apparaissaient dans la lettre de mission. La notion d'indicateurs de performance, mesurant l'efficacité et l'efficience des ressources électroniques c'est-à-dire, à la fois, l'usage des ressources électroniques gratuites ou payantes et de certains services afférents ainsi que leur efficience, a été retenue. Cette notion d'efficience est difficile à mesurer. En effet, le coût des ressources électroniques n'est pas toujours connu réellement car il est souvent la résultante d'un coût papier complété par un surcoût électronique. Il a également été décidé de suivre le plus possible les normes ISO en ce qui concernait le choix des items à retenir. La notion de ressources électroniques a été précisée, les définitions de la norme ne correspondant pas tout à fait à la réalité d'aujourd'hui ; on peut noter par exemple qu'il existe de plus en plus de périodiques électroniques sous forme de bases de données, il ne semble donc pas judicieux de repérer celles-ci indépendamment des périodiques, comme le souligne la norme NISO Z39.7 de 2004.

L'utilisation de l'expression document électronique (numérique) ou numérisé a été affinée. Une notion nouvelle a été introduite même si elle n'est pas encore très présente dans les établissements de l'enseignement supérieur. Il s'agit des services de référence bibliographique en ligne. Le Comité n'a pas retenu un certain nombre de mesures apparaissant dans les normes ISO et concernant les durées des sessions et les postes de travail, considérant que ces éléments étaient de la responsabilité des Services informatiques des universités. Quatre niveaux d'utilisateurs potentiels des résultats ont également été retenus : national (pour aider à l'évaluation dans le cadre de la LOLF), établissement (résultats indispensables au pilotage de l'établissement

et à l'évaluation de la politique documentaire), SCD (idem au niveau service) et Couperin (aide dans le cadre des négociations avec les fournisseurs). Enfin, le Comité a défini, après une longue discussion, les catégories d'usagers à identifier (population à desservir) concernant l'utilisation des ressources et les coûts afférents. La LOLF demande effectivement des données précises concernant les publics.

Un tableau final (Voir en annexe 9) a ainsi été élaboré et a été validé par les membres du Comité de suivi.

Il est à noter que ce tableau privilégie le quantitatif, une seule mesure concerne le qualitatif. Il faudrait peut-être étoffer ce dernier aspect si on en juge, par exemple, par l'évolution des travaux de l'ARL (Voir ci-dessus). De même, ce tableau pourrait être complété, en ce qui concerne les revues commerciales et en particulier pour le domaine STM, par des informations sur les performances desdites revues (l'« impact factor » tel que la société Thomson le définit et l'utilise dans le *Journal Citation Reports*).

Lors d'une intervention récente, P. Shepherd a parlé,, en alternative au facteur d'impact, de facteur d'usage (usage factor) ; faut-il introduire cette notion ?

Les items retenus n'ont été sélectionnés que par les membres du Comité de suivi. Le temps imparti à cette mission n'a pas permis d'effectuer une consultation plus large. Avant de pouvoir utiliser le tableau final, il paraît nécessaire d'en vérifier la faisabilité technique, quel que soit le mode d'obtention des données. Ce point sera développé dans le troisième chapitre du rapport.

Sur le plan de la construction des indicateurs, plusieurs articles (BIX, INRIA) soulignent la nécessité de faire appel à des statisticiens pour élaborer les modèles, vérifier la fiabilité des indicateurs… Une telle consultation s'est révélée indispensable lors de l'analyse des résultats des tests...

Les différents échanges en Comité de suivi ont également révélé la nécessité de proposer pour chaque item des définitions claires à communiquer aux établissements. La littérature insiste beaucoup sur ce point, montrant l'intérêt d'utiliser un même cahier des charges de récolte des données pour obtenir des résultats homogènes et donc comparables. A partir des définitions trouvées dans la littérature et particulièrement dans les normes ISO, un tableau de définitions (Voir en annexe 9), parallèle au tableau des items retenus, a été dressé.

Chapitre 2. Dispositifs techniques

Il s'agit ici, comme le demande la lettre de mission, de définir quels dispositifs mettre en place pour permettre aux établissements de collecter les données nécessaires pour remplir le tableau présenté dans le chapitre précédent. L'introduction de ce rapport précise déjà que le processus est différent selon qu'il s'agit d'exploiter des données statistiques externes ou de collecter également les données primaires.

Après l'étude bibliographique, des tests ont été lancés afin d'étudier la faisabilité de mesurer localement des statistiques sur l'usage des ressources électroniques. En parallèle, une analyse des statistiques préparées par les fournisseurs a été menée.

Enfin, des outils d'exploitation des résultats ont été étudiés et une synthèse en est présentée dans les pages qui suivent.

▪ Repères bibliographiques

Si de nombreux articles ont été écrits sur l'intérêt et la faisabilité d'effectuer

des mesures locales pour évaluer l'usage des ressources électroniques à l'aide de statistiques et d'indicateurs (T. Plum, B. Franklin…), peu d'articles concernent l'expérimentation elle-même, c'est-à-dire les processus et outils de mesure ou les logiciels utilisés. Cependant, quelques informations importantes ont été glanées ici ou là : des fichiers journaux (fichiers logs)[14] sont constitués à partir de la capture des transactions puis analysés. Cette méthode semble, selon la littérature, être la plus utilisée, ce que souligne par exemple H.R. Jamali, pour qui l'analyse des « logs » est une méthode efficace[15]. Un serveur mandataire (proxy server) est utilisé pour collecter facilement les données mais sans cache[16] afin de récupérer toutes les transactions sur le proxy et avoir une meilleure image du travail effectué ; une comparaison avec les données collectées par les fournisseurs est également possible. Ces informations ont servi de base pour la préparation des tests.

Le Code COUNTER donne également des recommandations dans ses directives (annexe D du Code jointe en annexe 10 de ce rapport) pour dresser les rapports statistiques. Dans les tests réalisés au cours de l'étude, les mêmes recommandations ont été suivies afin de pouvoir comparer de façon fiable les résultats obtenus localement avec ceux des fournisseurs, tout au moins pour ceux qui sont compatibles COUNTER. La comparaison entre données statistiques recueillies localement et données fournies par les éditeurs – et surtout les éventuels écarts qui seraient constatés – pourraient produire des résultats utilisables lors des négociations avec les éditeurs, notamment lorsque ceux-ci construisent leur modèle tarifaire sur l'usage que

[14] Voir glossaire en annexe 6

[15] Log analysis is an efficient evidence-based method for evaluation of the performance of a system such as a digital journal library against its objectives (The use... p. 558)

[16] Il reste le cache du PC, mais on peut considérer que dans ce cas, ce que l'usager effectue en plusieurs fois pourrait être effectué en une seule fois.

les établissements font de leurs ressources.

Il existe un certain nombre de logiciels de mesure des données transportées sur les réseaux entre l'utilisateur et le serveur de ressources, mais il s'agit généralement de logiciels à installer localement (Xiti, Awstats…) c'est-à-dire de logiciels utilisés pour mesurer les flux depuis l'extérieur vers un serveur interne et non pour mesurer des requêtes sur des serveurs externes et surtout des serveurs d'éditeurs commerciaux pour lesquels les logiciels utilisés ne sont pas connus ; le cas étudié dans ce rapport est donc plus difficile à traiter.

Par ailleurs, dans le cas présent, il s'agit d'extraire et d'analyser des masses importantes de données. Les dispositifs qui semblent nécessaires pour ce type de situation rejoignent des techniques d'exploration des données qu'on retrouve dans le monde économique, les banques… Il s'agit de techniques d'analyse de données et d'aide à la décision appelées « data mining »[17] et plus particulièrement dans ce cas « Web mining ».

Il y a peu de littérature sur les outils d'exploitation des résultats, qu'il s'agisse de l'extraction des données ou de traitement des résultats, à l'exception des informations commerciales. On voit cependant l'intérêt augmenter pour ces questions, parallèlement au développement des outils, ce que démontrent les différentes informations récoltées durant la mission. Le premier logiciel de gestion des ressources électroniques date de 2003[18].

La littérature souligne l'importance de disposer de tels outils d'extraction ou

[17] « Processus mis en œuvre pour faire émerger de la connaissance structurée à partir de données brutes, en se basant notamment sur des techniques d'analyse statistique ; ces connaissances permettront aux acteurs économiques de mieux comprendre leur domaine, et peut-être de prendre des décisions plus rationnelles » (S. Claudel, p. 31)

[18] On peut consulter l'article de S. Meyer (Helping…), qui recense fin 2005 un certain nombre de ces systèmes et cite Innovative comme premier éditeur

de traitement ; dans le cas contraire, le travail est très long puisqu'il faut gérer chaque éditeur séparément ; de plus toutes ces données sont souvent hétérogènes.

Pour compléter l'étude bibliographique, des contacts intéressants ont été pris soit avec des fournisseurs, pour leur demander des précisions sur leur mode de collecte de données statistiques, soit avec le directeur de projet de COUNTER, pour affiner les définitions du Code de bonnes pratiques, pas toujours claires, soit encore avec J.C. Bertot et R. Poll (cf annexe 8) pour partager leurs expériences. Des contacts avec certains fournisseurs de logiciels d'exploitation des résultats ont permis d'analyser leurs produits à travers des démonstrations.

Une des difficultés rencontrées dans cette étude, et par là même sans doute une de ses limites, a été d'obtenir des réponses claires aux diverses questions adressées aux uns et aux autres.

- **Mesures locales**

A partir d'une série de tests réalisés avec plusieurs établissements (Voir introduction), l'étude a cherché à mesurer localement l'usage des ressources électroniques à partir de l'analyse des flux de données transitant des réseaux locaux des établissements vers les fournisseurs, via l'internet.

Selon la littérature, plusieurs étapes sont nécessaires avant d'obtenir des résultats :

- la collecte des données avec la construction de fichiers journaux (fichiers « logs »),

- l'extraction parmi les journaux collectés de ceux qui sont utiles pour les mesures,
- l'exploitation des données en fonction des items retenus,
- la mise en forme des résultats.

Ce processus a été suivi pendant les tests. Pour la clarté du rapport, les deux premières étapes sont décrites ensemble, puis la troisième. La dernière étape est présentée dans la partie consacrée à l'exploitation des résultats.

Comme cela a déjà été mentionné, la première difficulté a été de mettre en place des cellules de tests avec des informaticiens. Cette collaboration a été recherchée parmi les établissements membres du Comité de suivi et également avec Grenoble, établissement ayant déjà une expérience intéressante. Ce choix ne reflète pas forcément la situation nationale.

Les premières étapes concernent la collecte et l'extraction des données[19]. Un serveur mandataire est utilisé, à savoir un « proxy web», comme outil de collecte des fichiers journaux. Il supporte, entre autres, les accès à des bases construites selon le protocole http(s). Ce serveur peut être réservé à la documentation ou servir à tout l'établissement. Tous les utilisateurs doivent passer obligatoirement par cette machine qui stocke les fichiers journaux correspondant au flux de requêtes vers tel ou tel fournisseur.

Lorsque les serveurs mandataires sont gérés par les universités, les flux transitant par ces machines ne concernent pas que la documentation électronique et génèrent donc des flux d'information importants. Un filtre

[19] Une documentation technique, décrivant l'expérience de Lyon 1 est présentée en annexe 11

sélectionnant chaque nuit les seules données à conserver, à partir d'une suite d'URLs pré-établie[20] caractérisant les plates-formes des éditeurs auxquelles on souhaite accéder, est nécessaire car une fraction minime doit être conservée pour les mesures (Lyon 1 évalue à 14% soit 466 Mo la part concernant la documentation et sur ces 466 Mo, 19% environ soit 90 Mo servent chaque mois pour effectuer les statistiques). La fonction cache du serveur proxy doit être annihilée (voir ci-dessus) ; les proxies n'étant pas toujours gérés directement par les SCD, ce point doit être négocié avec les universités pour la documentation électronique.

De ces premiers essais, il ressort que le paramétrage des proxies, s'il s'agit des machines générales de l'université et non de machines dédiées à la documentation électronique, doit parfois être modifié pour permettre la collecte des données attendues ; c'est en tout cas ce que les tests réalisés à l'UCBL et à l'UTC qui utilisent un serveur Squid[21] ont montré, les services informatiques n'ayant pas installé ces serveurs pour répondre à l'objectif de la documentation électronique. Le paramétrage des proxies comme les filtres installés ont suivi principalement les directives du Code COUNTER, ceci afin de pouvoir comparer les résultats avec ceux des fournisseurs. On peut citer les éléments suivants : pas de cache comme cela a déjà été mentionné, heure UTC/GMT afin de considérer une journée ou un mois de statistiques avec les mêmes critères, que les fournisseurs se trouvent sur les mêmes fuseaux horaires ou non, fonction « query string » pour récupérer le plus d'informations possible ; élimination de certains codes retour inutiles pour la collecte des données, gestion des « doubles clics » pour supprimer les clics

[20] Du type « acs.org » « aip.org » « aps.org » « asm.org »…
[21] Voir glossaire en annexe 6

intempestifs….

Il est à noter que tous les tests ont été faits sur des proxies utilisant le logiciel Squid ; il semblerait que le type de logiciel ne soit pas un frein pour les paramétrages retenus ; ce point n'a pu être vérifié concrètement.

Tous les éditeurs n'acceptent pas de recevoir une seule adresse IP (celle du proxy), situation de plus en plus rare aujourd'hui.

Pour réduire la taille des fichiers, il est utile de les nettoyer et de ne retenir que les données exploitables ; Lyon 1 évalue, en première approximation, à 1 Go par an le stockage nécessaire ou le double si on stocke l'année en cours et l'année précédente[22]. Effectivement, le Web génère une grande quantité d'informations et donc de transactions qui ne servent pas pour produire des statistiques. Seules, celles qui correspondent à des requêtes intentionnelles et réussies, servent aux mesures.

Exemple de transaction à conserver suite à la requête : téléchargement d'un article en pdf
1164956924.166 1296 XXXX TCP_MISS/200 26159 GET http://pubs.acs.org/cgi-bin/article.cgi/jacsat/2003/125/i33/pdf/ja035175y.pdf -DIRECT/216.143.112.80 application/pdf

Cependant, les requêtes échouées sont intéressantes car elles permettent d'effectuer d'autres contrôles : savoir, par exemple, quels titres ont été demandés mais sans succès, par manque de droits d'accès.

Le paramétrage du proxy comme le filtrage des données sont des étapes essentielles car elles augurent de tous les résultats et de comparaisons

[22] Il est cependant nécessaire de conserver au moins deux mois les données brutes afin de pouvoir, si les statistiques sont faites chaque mois, contrôler les URLs filtrées et les modifier éventuellement au vu des résultats (pour Lyon1, cela correspondrait à environ 1x1 Go).

ultérieures fiables avec les rapports des fournisseurs.

On peut se demander si la solution serveur mandataire est envisageable dans les établissements universitaires français. Dans l'enquête mentionnée en introduction, à laquelle 64 établissements sur 102 interrogés ont répondu, on constate que 50% environ des établissements ont un proxy, que ce soit au niveau du SCD ou du CRI de l'Université.

Les dernières étapes concernent l'exploitation des données et la préparation des résultats.

Les tests ont été effectués d'abord sur les revues. La raison de ce choix est que ce sont les documents électroniques les plus utilisés à ce jour ; compte tenu des budgets concernés dans les établissements, étudier la faisabilité d'obtenir toutes les statistiques nécessaires à la gestion de ces documents ou d'exercer un contre-pouvoir lors des négociations avec les fournisseurs par le biais de statistiques internes a paru prioritaire. Des tests de mesure d'usage des bases de données bibliographiques et encyclopédies ont également été effectués dans un deuxième temps. Seul l'usage des monographies n'a pu être étudié par manque de temps et de moyens humains.

Les tests ont été limités à quelques-uns des items retenus par le Comité de suivi, ils ont porté essentiellement sur le nombre de recherches, le nombre d'unités documentaires téléchargées concernant les résumés, les sommaires, les articles en pdf ou en html, les adresses IP... (Voir annexe13).

Deux outils avec des technologies différentes ont été retenus pour l'étude :

tous les deux utilisent des scripts en PERL[23] (Practical Extraction and Report Language), mais Compiègne a recours également à un paramétrage et une sortie en XML (Extensible Markup Language) alors que Lyon et Grenoble n'utilisent que du PERL[24]. A partir des fichiers journaux, des compteurs sont ainsi créés par item mesuré.

```
# format:
# http://pubs3.acs.org/acs/journals/toc.page?incoden=jpcafh
# http://pubs.acs.org/subscribe/journals/ancham/jtoc.cgi?ancham/last/last

$regexp_TITRE_TOC_1 = '^http://pubs[^\.]*\.acs\.org/acs/journals/toc\.page\\?incoden=([a-z1-9]+)';
$regexp_TITRE_TOC_2 = '^http://pubs[^\.]*\acs\.org/.*/jtoc.cgi';
$regexp_TITRE_TOC = "$regexp_TITRE_TOC_1|$regexp_TITRE_TOC_2";
```

Des tableaux finaux de résultats sont élaborés[25] :

- Par titre et date

T_ABST_HTM	T_ABST_PDF	T_TOC_HTM	T_FULL_PS	T_FULL_HTM	T_FULL_PDF	T_ARTICLE_PDF	T_SAMPLE_PDF	T_ASAP_PDF	T_ARCHIVE_PDF	T_SUPPINFO_PDF	EDITEUR	ANNEES	TITRES
20	16	88		6		58	2	12	18		acs.org	TOTAL	Accounts of Chemical Research,
2						8	4				acs.org	2006	ACS Chemical Biology,
8						6	2				acs.org	2007	ACS Chemical Biology,
10		14				14	6	4			acs.org	TOTAL	ACS Chemical Biology,
	8										acs.org	1950	Analytical Chemistry,
	6								4		acs.org	1952	Analytical Chemistry,
.....													
				2	64						rsc.org	2007	Chemical Communications
				56	328						rsc.org	TOTAL	Chemical Communications

[23] Voir glossaire en annexe 6
[24] Voir les détails dans la documentation technique, en annexe 11
[25] Les données sont fictives

- Par éditeur

TOT_PS	E_SESSION_HTML	_URNAWAY_HTML	E_SEARCH_HTM	E_ABST_HTM	E_ABST_PDF	E_TOC_HTM	E_FULL_PS	E_FULL_HTM	E_ARTICLE_HTM	E_SAMPLE_HTM	E_ASAP_HTM	E_FULL_PDF	E_ARTICLE_PDF	E_SAMPLE_PDF	E_ASAP_PDF	E_ARCHIVE_PDF	E_SUPPINFO_PDF	E_TELECH_HTML	PLATEFORMES
			3834	4796	710	7266			786	48	130		8846	374	1780	2698	24		acs.org
26			556	240		820		86				2244							aip.org
																			aps.org
			1170	3030		2040		392				3730							blackwell-synergy.com
			5182	150		3520		6316				3028							emc-consulte.com
			3524	6858		1240		266				7568							interscience.wiley.com
																			iop.org
	4718																	1588	isiknowledge.com
				2752		2802		204				2854							rsc.org
																			sagepub.com
			34688	6870		10128		46				44180							sciencedirect.com
			334	406		330		640				706							sciencemag.org
		6	350	54														4162	silverplatter.com

Une comparaison de ces outils a été amorcée. Au point de la réflexion, il semblerait préférable de privilégier des scripts tout en PERL et des tableaux de résultats structurés en excel ou XML[26] afin que les bibliothécaires puissent plus facilement exploiter les résultats.

Une première analyse des fichiers journaux montre que le repérage des informations indiquant une recherche sur tel titre, un téléchargement pdf, html… n'est pas toujours évident et demande d'effectuer des simulations fines du travail accompli par les utilisateurs potentiels afin d'identifier les chaînes de caractères servant à constituer les compteurs. Par ailleurs, les données sont souvent codées dans les fichiers journaux et les éditeurs sont parfois les seuls à pouvoir fournir certaines informations.

Une mesure intéressante concerne l'usage par titre de revue. L'identification des titres dans les fichiers logs n'est pas toujours l'ISSN ou le Coden et peut donc être difficile à transcrire. Il serait utile que les fournisseurs donnent leurs tables de correspondance. Ce point devrait être précisé dans les contrats

[26] Les tableaux présentés sont des tableaux excel.

négociés avec les éditeurs ou agrégateurs.

De même, les tests ont permis de compter par date, pour un titre donné, les éléments téléchargés. Cette mesure peut être utile aux établissements pour leur gestion des archives et a été rajoutée récemment dans le tableau des items.

Par ailleurs, les unités de contenu documentaire téléchargées par page (Voir tableau des items retenus en annexe 9), à savoir les éléments « sommaire, résumé, bibliographie, requêtes sur le texte intégral en pdf-html-ps, refus de connexion pour le texte intégral en pdf-html-ps »[27] proposés dans le rapport JR3 des directives COUNTER, ne peuvent pas toujours être obtenues de façon unique, car si la même page html propose différents éléments, le code dans le « fichier log » est le même et on ne peut alors dissocier par exemple bibliographie et résumé.

Les mesures locales permettent d'affiner les résultats sur les unités de contenu documentaire téléchargées par rapport à ceux que donnent les fournisseurs. On peut ainsi repérer, au moins pour certains éditeurs, archives et articles courants[28], articles payants ou gratuits… (Voir tableau p. 33).

Outre les compteurs cités ci-dessus, d'autres ont également été étudiés. Concernant le nombre de sessions[29], la mesure n'a pu être testée à fond. Des tests sur les sessions rejetées, dans le cas de licences calculées au nombre d'utilisateurs simultanés, ont été effectués pour un seul fournisseur de bibliographie. Parmi tous les éditeurs testés, le nombre de sessions a pu

[27] "TOC, abstract, references, full-text pdf-html-ps, turnaways full-text pdf-html" selon le Code COUNTER

[28] Depuis la réalisation de cette mesure, un nouveau rapport JR1a non obligatoire concernant les articles d'archives a été ajouté par COUNTER, preuve de l'intérêt du test effectué ; ce rapport n'est pas encore produit par les fournisseurs.

[29] Voir annexe 9

être calculé une seule fois à partir des fichiers journaux. La littérature souligne la difficulté d'avoir aujourd'hui une interprétation homogène de la notion de session, tout particulièrement depuis l'arrivée des bases Web.

Le Comité de suivi souhaite des statistiques par catégorie d'utilisateurs. Des résultats par adresse IP sont obtenus (ce test a été effectué à Compiègne, le processus fonctionne déjà à Grenoble). Si les établissements possèdent des tables de correspondance IP/catégorie d'usagers, des résultats par catégorie sont possibles. Il semble que les établissements ne gèrent pas toujours leurs adresses IP selon ce critère, c'est d'ailleurs souvent le cas dans les bibliothèques. Ces statistiques peuvent donner des résultats très intéressants tout en respectant les règles de confidentialité édictées par la CNIL ; il suffit d'« anonymiser » les IP.

Le problème est plus compliqué lorsqu'il s'agit d'utilisateurs nomades.

Une solution serait une combinaison proxy/CAS (Central Authentication Service) ou encore proxy/annuaire LDAP (Lightweight Directory Access Protocol). Le logiciel de paramétrage Squid accepte les identifications par login/mot de passe, ce qui, combiné à un annuaire ou un CAS, permettrait sans doute d'effectuer des mesures par catégorie d'usagers. Tout ceci peut être combiné à une interrogation nomade via un VPN (Virtual Private Network).

Shibboleth[30], protocole pour lequel une expérience menée par le CRU vient de s'achever, permet d'effectuer un contrôle d'accès ciblé pour chaque usager nomade[31], contrôle effectué par le fournisseur auprès de

[30] Voir glossaire en annexe 6

[31] Shibboleth peut catégoriser finement les utilisateurs mais cette fonctionnalité n'a pas été retenue

l'établissement, l'usager se connectant directement. Cependant, le CRU a précisé que l'architecture de Shibboleth n'était pas compatible aujourd'hui avec la présence d'un proxy. Les établissements souhaitant obtenir des statistiques locales détaillées ne pourront donc utiliser Shibboleth pour la documentation électronique, ce protocole permettant de savoir uniquement quels fournisseurs ont été interrogés. Ces solutions n'ont pu être testées faute de temps, l'expérience Shibboleth s'étant achevée en décembre 2006.

Tout ce travail sur les scripts doit bien sûr être fait pour chaque éditeur car ceux-ci n'utilisent pas toujours les mêmes termes pour exprimer « recherche, affichage »… ou pas toujours avec le même sens. Des mises à jour sont également nécessaires chaque fois que les fournisseurs modifient leurs logiciels, ce qui risque d'alourdir le processus. Durant le temps passé aux tests pendant la mission, plusieurs scripts ont dû être modifiés suite à des changements du côté du fournisseur.

Les tests effectués permettent d'induire le temps nécessaire à la création de nouveaux compteurs : 20 heures pour un éditeur donnant lieu à cinq compteurs. En ce qui concerne le suivi, le temps peut varier de quelques heures par mois s'il s'agit de faire passer les scripts et éventuellement de modifier le contenu d'un compteur à 20 heures si un éditeur a totalement changé de plate-forme.

Des tests sur des données libres de droit du type archives ouvertes avaient été envisagés mais n'ont été réalisés qu'à Compiègne et sur la base HAL

(Centre pour la communication scientifique directe-CCSD). En effet, les mesures locales imposent le passage par un proxy, ce qui est difficile à exiger dans le cas de ressources libres.

Malgré un certain nombre de limites ou contraintes mentionnées ci-dessus, des résultats intéressants, principalement en ce qui concerne les revues, ont été obtenus. Un tableau des 13 éditeurs testés à Lyon 1 est joint en annexe 12, d'autres compteurs existent à Grenoble, ce qui donne un total de 25 éditeurs étudiés[32]. Le tableau montre quels compteurs ont pu être mis en place, il dresse un comparatif avec les mesures des éditeurs conformes ou non au Code COUNTER ; ce tableau[33] souligne également que les mêmes résultats n'ont pu être obtenus pour tous les éditeurs.

Pour pouvoir connaître la faisabilité des mesures locales, il est important de les comparer dans un premier temps aux résultats des fournisseurs, lorsqu'ils sont connus. A cet effet, une analyse statistique des résultats des compteurs testés à Lyon 1 a été effectuée par le stagiaire (Lyon 1) et est présentée de façon détaillée en complément à l'annexe 12. Le coefficient de corrélation linéaire permet de vérifier si les résultats locaux et ceux des fournisseurs évoluent de façon similaire ; si oui, on vérifie que les données locales sont proches de celles des éditeurs à l'aide du principe de régression linéaire[34]. L'interprétation de ces résultats numériques permet d'en tirer quelques conclusions.

Les compteurs pour les revues donnent des résultats globalement satisfaisants et permettent d'évaluer correctement l'usage de ce type de

[32] Les éditeurs étudiés à Compiègne et Strasbourg ne sont pas comptabilisés ici.
[33] Classé par type de document
[34] Cette analyse n'a pu être conduite de façon satisfaisante pour tous les éditeurs étudiés par manque de résultats.

document. Parfois, les résultats locaux divergent des rapports des fournisseurs, mais il n'est pas sûr que ces derniers soient toujours exacts.

En ce qui concerne les encyclopédies, les compteurs mis en place semblent pouvoir permettre d'effectuer une bonne évaluation de leur usage mais, à ce jour, les données récoltées sont insuffisantes pour conclure en ce sens.

Enfin, peu de bibliographies ont pu être testées. Le faible nombre de compteurs mis en place et le peu de mesures disponibles à ce jour ne permettent pas de valider la faisabilité des mesures locales pour ce type de document.

- **Données externes produites par les fournisseurs**

L'obtention de statistiques étant essentielle pour les établissements, ceux-ci doivent s'assurer que les fournisseurs leur procurent les données qui leur sont nécessaires, s'ils ne souhaitent pas mettre en place un système de mesures locales.

Les fournisseurs peuvent donner des statistiques conformes ou non au Code COUNTER.

L'enquête d'août 2006, déjà citée, a permis de repérer quelles statistiques les établissements récupéraient. Même si la période n'était pas propice aux enquêtes, les résultats étant conformes aux informations récoltées auprès des bibliothèques membres du Comité de suivi interrogées au printemps 2006, il semble que les données soient fiables.

Au moment du dépouillement de l'enquête, sur 144 éditeurs, agrégateurs ou plates-formes de revues et bases de données présents dans les établissements, 37 sont déclarés compatibles COUNTER (version 2) sur 55

présents dans la liste COUNTER à cette date, une dizaine ne sont représentés que dans 1 ou 2 SCD. 36 éditeurs soit 24% ne fournissent aucune statistique. Il y a donc plus de rapports statistiques « non-COUNTER» (71 fournisseurs soit 62% de ceux qui produisent des statistiques) que de rapports COUNTER. Parmi ces éditeurs, huit d'entre eux (7%) fournissent des rapports du type COUNTER sans être homologués (il s'agit en partie d'éditeurs qui étaient conformes à la version 1 du Code mais qui n'ont pas évolué).

En ce qui concerne les livres et autres ouvrages de référence, seuls trois éditeurs fournissaient en novembre 2006 des statistiques compatibles COUNTER[35]. Lors d'un exposé en janvier 2007 à la Conférence de l'ICSTI (International Council for Scientific and Technical Information), P. Shepherd soulignait la difficulté de définir des mesures d'usage pour les livres aussi pertinentes que celles qui existent pour les revues[36].

La difficulté est de savoir si les rapports fournis respectent bien les directives COUNTER. A partir de l'exemple d'un établissement, on constate que tous les éditeurs ne fournissent pas une documentation explicative (la moitié pour le SCD pris en exemple), et lorsqu'ils en fournissent une, ils se contentent souvent de renvoyer vers le Code de bonnes pratiques. Dans la documentation d'un éditeur, il est clairement écrit, fin 2006, qu'il ne respecte pas les directives COUNTER alors qu'il est mentionné comme étant conforme. Il n'est pas aisé de contrôler la fiabilité des rapports ; on a cependant pu vérifier, par exemple, que tous les éditeurs ne respectaient pas

[35] Voir http://www.projectcounter.org/News_release_November_2006.doc
[36] "[For books], relevant usage metrics less clear than for journals", vue 4

le délai de fourniture desdits rapports. Ils n'envoient pas toujours, comme le recommande le Code, les rapports aux établissements d'une part, ceux qui sont destinés aux consortia d'autre part. Les premiers tests laissent également planer quelques doutes sur l'exactitude des données, ce que souligne un SCD dans l'enquête en écrivant, concernant les rapports, « il nous arrive de douter de leur fiabilité ». Les fournisseurs semblent ne pas toujours bien gérer leurs fichiers, ce qui peut fausser les statistiques[37]. C'est ainsi que le consortium Couperin s'est aperçu que des établissements ne faisant pas partie du groupement de commandes apparaissaient néanmoins dans les statistiques fournies en l'occurrence par certains éditeurs. Un travail manuel important est alors nécessaire pour obtenir des résultats fiables.

Un certain nombre d'éditeurs signalent régulièrement qu'ils doivent calculer à nouveau leurs statistiques, celles-ci s'étant avéré erronées ; n'y a-t-il pas des erreurs non détectées ? Selon la littérature (P.M. Davis), tant qu'il ne s'agit pas de la même plate-forme et du même logiciel d'application, on ne peut jamais être certain d'obtenir des statistiques comparables : par exemple un fournisseur impose pour obtenir un document en pdf de télécharger d'abord le document en html, un autre n'aura pas cette contrainte…

Un certain nombre d'éditeurs fournissent des rapports complémentaires, en particulier sur les titres auxquels les établissements ne sont pas abonnés mais pour lesquels les utilisateurs ont essayé de décharger des articles (Wiley…). Il est dommage que le Code COUNTER n'impose pas cette donnée car elle est essentielle pour les établissements dans l'élaboration de

[37] Voir par exemple le message concernant OUP http://www.babouin.fr/post/2007/01/22/Probleme-sur-les-statistiques-doxford-University-Press-pour-2006

leur politique documentaire. De même, le rapport COUNTER JR3 (Voir annexe 7) qui pourrait aider les bibliothèques à comprendre comment les utilisateurs interrogent les ressources (par consultation des sommaires ou par recherche…), ce qui serait intéressant pour les formateurs dans les bibliothèques, ne fait pas partie de la liste des rapports obligatoires. Un SCD déplore que « certains éditeurs n'aient pas mis en place tous les indicateurs ».

Une des incertitudes qu'on peut relever, dans les rapports COUNTER mais sans doute aussi chez les autres fournisseurs, est due au manque de clarté entre les statistiques des éditeurs et celles des agrégateurs. C'est ce que soulignent A. Conyers et P. Dalton dans le rapport sur les statistiques d'usage réalisé dans le cadre de l'initiative NESLi2[38].

Si on résume l'avis des établissements concernant les statistiques en leur possession, les rapports COUNTER satisfont globalement les utilisateurs : « dans le cas d'abonnements souscrits depuis plusieurs années, auprès de sociétés compatibles Counter (Elsevier entre autres), les données recueillies sont fiables et constituent des éléments de réflexion particulièrement intéressants, notamment dans le cadre du financement de la documentation numérique », note un SCD dans l'enquête. Certains regrettent cependant des insuffisances : changement de plate-forme d'une année sur l'autre, ce qui donne des résultats non comparables, définition des données peu claires, fusion d'éditeurs, et tout particulièrement manque de statistiques par adresse IP…

[38] Voir glossaire en annexe 6

En ce qui concerne les fournisseurs « non COUNTER », et selon l'enquête, les statistiques ne sont pas toujours présentes, pas toujours exploitables, pas toujours celles que l'on souhaite. Elles se limitent parfois au nombre de connexions. Elles comportent peu de résultats sur l'activité des usagers (IP…). Par contre, les établissements les obtiennent directement, ce qui est sans doute plus simple.

Selon l'enquête déjà citée, les résultats, COUNTER ou non, sont plus fiables pour les revues que pour les bases de données qui ne fournissent souvent pas de statistiques via une interface Web, celles-ci étant alors peu exploitables. Il est difficile, lorsqu'il s'agit de plates-formes gérant un ensemble de bases, d'extraire les données correspondant à chaque base lorsque les recherches sont multi-bases. « [Les statistiques] ne sont pas adaptées lorsque plusieurs bases sont regroupées sous une même interface : on ne sait pas si une interrogation de trois bases simultanément compte pour trois interrogations ou une seule. Qu'est-ce qui est pris en compte si l'on entre d'abord dans une base puis si l'on passe de celle-ci à une autre base de la même interface ? » s'interroge une bibliothèque.

Les établissements à dominante scientifique sont plus satisfaits par les rapports des fournisseurs que les autres, ce qui s'explique par le fait qu'ils obtiennent globalement plus de rapports COUNTER et donc des rapports plus homogènes. Les établissements à dominante droit/économie et SHS ont plus de ressources provenant d'éditeurs français, absents, comme cela a déjà été souligné, de COUNTER. C'est aussi dans ces disciplines et en particulier en droit/économie, que les statistiques manquent le plus.

- **Exploitation des résultats**

Quel que soit le mode d'obtention des données statistiques, il est également nécessaire de les exploiter.

Dans le cas de statistiques locales, les outils qui ont été utilisés dans les tests produisent des rapports excel. Il serait intéressant de présenter les résultats structurés selon le format XML afin de faciliter leur exploitation par les bibliothécaires. COUNTER le propose dans sa 2ème édition pour les revues et les bases de données.

A partir des données statistiques, il est possible de créer les indicateurs attendus en combinant différentes données.

L'intérêt des statistiques locales est que les résultats sont homogènes et obtenus directement, « l'intérêt est d'avoir des statistiques uniformes, quel que soit l'éditeur. Cela nous permet également de comparer nos résultats avec les statistiques fournies par les éditeurs », précise un SCD.

Dans le cas des statistiques des fournisseurs compatibles COUNTER, leur récupération est longue puisqu'il faut aller sur chaque site ; parfois il faut lancer les rapports, ce qui peut prendre du temps, de plus, ils ne sont pas toujours disponibles à la même date. Les tableaux récupérés étant parfois différents d'un fournisseur à l'autre, une harmonisation peut s'avérer nécessaire.

L'initiative SUSHI (Standardized Usage Statistics Harvesting Initiative) créée sous l'égide de NISO, est un projet de norme Z39.93 (Draft standard) proposant un protocole unique en XML qui permet aux établissements de

collecter automatiquement les rapports statistiques d'usage conformes aux directives COUNTER et d'analyser lesdites statistiques depuis leur système de gestion. Mais, n'intégrant que des données conformes au Code COUNTER, ce protocole ne donne pas à un établissement la possibilité de gérer directement toutes les statistiques de ses ressources électroniques. Ce protocole est toujours au stade expérimental[39].

Les données étant récupérées, il faut les exploiter, dresser des tableaux… Des systèmes d'ERM (Electronic Resource Management System)[40] existent ; ils ont l'intérêt de proposer un ensemble de fonctionnalités de gestion (abonnements ou acquisitions, statistiques…). La plupart gèrent les données des fournisseurs compatibles COUNTER, mais pas seulement. Ces systèmes intègrent, semble-t-il, au moins pour partie, le modèle SUSHI. On peut citer, à titre d'exemple[41], la société Innovative qui a intégré ce modèle dans son ERMS « Innovative Electronic Resource Management » ou la société Ex Libris qui l'a testé dans une université avec son ERMS Verde.

Les ERMS sont des outils a priori intéressants ; grâce à eux, un établissement peut obtenir, entre autres, les éléments nécessaires à la constitution de ses indicateurs dans un ensemble intégré. Il n'est pas sûr que tous les items attendus (Voir premier chapitre) puissent être calculés.

Certains ERMS sont vendus séparément des autres modules de gestion (bases de connaissance, SIGB…)[42]. Dans le cas de systèmes hétérogènes, le problème d'interfaçage doit être analysé finement même si le fournisseur

[39] Information d'Ebsco
[40] Voir glossaire en annexe 6
[41] Seuls, sont cités les outils qui ont pu faire l'objet d'une présentation, il est dommage que ces deux sociétés ne fournissent pas de version de test. Voir aussi la bibliographie qui présente d'autres outils.
[42] C'est le cas d'Innovative et d'Ex Libris.

l'annonce comme possible.

A été également créé par MPS Technologies le service ScholarlyStats qui s'appuie sur le modèle SUSHI, afin de fournir aux professionnels de l'information un outil permettant d'analyser l'utilisation de leur contenu en ligne. Des rapports de statistiques d'utilisation sont collectés et traités tous les mois par MPS Technologies, mais manuellement, à partir des données brutes des fournisseurs de périodiques ou de bases de données afin de les uniformiser en un seul rapport standardisé. Au dire des responsables, ces rapports sont accessibles aux professionnels de l'information à travers une interface « conviviale ». Les responsables de Scholarlystats ne fournissent que les rapports de la version 1 de COUNTER. Il faut noter que, si MPS Technologies s'appuie sur le modèle SUSHI, il intègre également des plates-formes « non-COUNTER » (11 sur les 42 traités à ce jour) ; 31 sur 55 éditeurs COUNTER sont ainsi répertoriés. A ce jour, aucun éditeur français ni même allemand n'est présent[43]. Ex Libris prépare l'intégration de Scholarlystats dans son ERMS pour le printemps 2007.

La société Thomson vient de commercialiser un nouvel outil JUR (Journal Use Reports). L'intérêt de cet outil est que, dans le même ensemble, on peut connaître l'usage de telle revue (déchargement) sur telle plate-forme de tel éditeur compatible COUNTER, mais aussi connaître pour l'établissement l'activité de citation… ; éléments qui pourraient sans doute être intégrés dans des modèles économiques. Thomson travaille avec MPS Technologies sur Scholarlystats.

L'émergence de tous ces outils permettant l'exploitation des résultats est

[43] Information obtenue en janvier 2007

intéressante. Cependant, les adaptations paraissent prendre du temps, ce qui semblerait prouver à la fois une attente forte de la part des établissements et une difficulté de mise au point pour de tels outils.

Dans le cas des fournisseurs « non-COUNTER », l'exploitation est encore plus difficile (Scholarlystats n'intègre aucun éditeur français) car les résultats par éditeur sont hétérogènes, obtenus, selon l'enquête d'août 2006, plus ou moins régulièrement, de façon souvent partielle ou même inexploitable. C'est pourquoi sans doute, les établissements n'arrivent-ils pas à traiter tous les résultats. Selon l'enquête, lorsqu'ils les exploitent, c'est la plupart du temps avec des tableurs excel.

Il est nécessaire, si on veut faire un suivi sur plusieurs années, de récupérer les données au fur et à mesure de leur parution. En effet, comme le souligne la littérature, il y a de tels changements (fusion d'éditeurs…) qu'il faut créer ses propres bases de données au fur et à mesure, et non compter sur les serveurs des éditeurs.

Chapitre 3. Synthèse

Après avoir défini quels items seraient pertinents pour permettre aux différents utilisateurs de statistiques et indicateurs (Ministère, établissements…) de récolter les informations qui leur sont nécessaires, après avoir analysé des modes de collecte de toutes ces données et des outils permettant de les exploiter, ce troisième chapitre essaie de dresser une synthèse ou plutôt un comparatif des solutions possibles à partir du modèle « statistiques locales » ou du modèle « COUNTER », sachant que certains fournisseurs non compatibles COUNTER produisent des rapports du type COUNTER sans être bien sûr « audités ». Dans ce comparatif, les autres rapports « non-COUNTER » ne sont pas pris en compte car, comme cela a été dit dans le chapitre précédent, ils sont très hétérogènes, souvent inexploitables ; il ne semble pas utile de les analyser davantage.

- **Comparaison items Comité de suivi / items COUNTER**

Un tableau mettant en regard les rapports statistiques que fournissent les éditeurs compatibles COUNTER avec les données retenues par le Comité de suivi est proposé en annexe.

Aucune statistique n'est fournie dans le Code COUNTER par catégorie d'usagers. Il est normal que ces catégories ne soient pas connues de façon précise par les éditeurs mais des résultats par adresse IP « anonymisée »[44] devraient être possibles puisque certains éditeurs compatibles COUNTER les fournissent déjà, hors rapports officiels. Par ailleurs, si le rapport JR3 de COUNTER était obligatoire, la compatibilité avec les attentes du Comité de suivi serait meilleure. Ce dernier a souhaité bénéficier de données permettant d'évaluer l'usage « intellectuel » des ressources, données que l'on retrouve dans ledit rapport.

Les rapports COUNTER ne fournissent pas de résultats sur les unités de contenu documentaire téléchargées pour les livres et les enregistrements téléchargés pour les bases de données bibliographiques.

En ce qui concerne les statistiques pour les consortia, le Code COUNTER a moins d'exigences que le consortium Couperin en matière de rapports.

- **Comparaison items Comité de suivi / solution locale**

Des tests effectués, qui ne concernent que certains éditeurs (Voir annexe 12)

[44] Dans ce cas, l'université doit connaître la correspondance IP/catégorie d'usagers que ce soit le type LMD, le nom d'un laboratoire…

et ne couvrent pas tous les domaines de la connaissance, il ressort que les compteurs pour les revues donnent des résultats globalement conformes aux attentes du Comité (Voir comparatif en annexe 13), plus complets dans certains cas (type d'articles téléchargés). Cependant, pour avoir des résultats par titre, donnée demandée par le Comité de suivi, il est parfois nécessaire que les éditeurs donnent le décodage qu'ils utilisent (Voir chapitre précédent). Les compteurs mis en place pour les encyclopédies sont moins complets que pour les revues mais correspondent à peu près aux attentes.

Par contre, pour les bibliographies testées, il ne semble pas possible, en tout cas pour les tests effectués, de mesurer le nombre d'enregistrements téléchargés, donnée souhaitée par le Comité de suivi mais également absente des rapports COUNTER (Voir ci-dessus). Par ailleurs, plus que pour les revues, les mêmes items n'ont pu être mesurés pour chaque éditeur.

- **Comparaison solution locale / solution COUNTER**

Si on consulte la littérature, on constate que les auteurs optent pour l'une ou l'autre des solutions.

Certains auteurs soulignent l'importance d'avoir :

- des statistiques indépendantes de celles fournies par les vendeurs, ceci afin de pouvoir contrôler leurs résultats souvent difficiles à interpréter (J. Duy et L. Vaughan) et pour avoir plus de détails (T. Plum) ;
- des statistiques plus fiables pour mieux maîtriser les négociations avec les fournisseurs (B. Franklin) ;
- des statistiques élaborées localement pour obtenir des données homogènes (O. Poncin). Pour résoudre cette difficulté, l'Ohiolink (Ohio

Library and Information Network)[45] a par exemple chargé sur ses serveurs les livres électroniques de fournisseurs commerciaux (solution coûteuse il est vrai) ;

- des statistiques comparables entre les résultats provenant des ressources payantes et de celles libres de droit (B. Franklin) ;

A contrario, J. Luther pense qu'il est impossible d'effectuer des statistiques localement si ce n'est pour mesurer le nombre de sessions ou analyser le type d'utilisateurs, arguant du fait que seul l'éditeur peut tracer l'activité une fois que l'utilisateur est sur son site.

D'autres considèrent que le rapport JR1 (COUNTER) est très satisfaisant et peut être considéré comme l'unité de mesure de l'usage (A. Conyers)[46].

Des contacts pris, il ressort également que :

- en ce qui concerne le choix de statistiques mesurées localement, J.C. Bertot, lors d'une rencontre durant la mission, a précisé que l'ARL avait décidé, après plusieurs projets (Voir premier chapitre), d'utiliser les statistiques des fournisseurs produites selon le Code COUNTER. J.C. Bertot explique ce choix par le fait que réaliser localement des mesures est une tâche trop lourde. L'ARL étant un des fondateurs de COUNTER, on peut en déduire l'intérêt de créer un standard pour la production de statistiques.

- l'Université de Münster utilise un système mixte mais prend les statistiques des fournisseurs compatibles COUNTER lorsqu'elles existent (R. Poll).

- même si certaines universités, membres de la Conférence collectent

[45] A of Ohio's college and university libraries and the state Library of Ohio

[46] "The study team … has assisted in the testing and validation of "the successful full-text article request" (COUNTER JR1 report) as a possible unit of measurement of "usage"…"

leurs propres statistiques (Ontario), le CREPUQ (Conférence des Recteurs et des Principaux des Universités du Québec) a finalement choisi d'utiliser les statistiques des fournisseurs respectant le standard COUNTER ; Madame Belzile, présidente du Sous-comité des bibliothèques, considère néanmoins que les questions soulevées en France sur l'intérêt d'avoir des statistiques indépendantes de celles des fournisseurs ou communes à celles des archives ouvertes sont intéressantes.

On peut se demander si, hormis le cas de l'Université de Münster, ces choix sont transposables en France. Selon les résultats de l'enquête (Voir chapitre précédent), seuls 26% des fournisseurs présents dans les établissements sont compatibles COUNTER, et aucun n'est français. Si on en juge par la solution choisie par l'Université de Münster, COUNTER n'est peut-être pas la solution unique, au moins aujourd'hui et en dehors des pays anglo-saxons.

L'analyse des deux solutions (mesures locales / rapports COUNTER)[47] conduit à faire quelques remarques. La solution locale assure l'homogénéité des résultats statistiques au sein d'un établissement mais aussi entre les établissements si le même outil est utilisé sur un plan national, ce qui offrirait une comparaison fiable au niveau national. Cette solution permet également aux établissements de créer des rapports divers en fonction de leurs propres besoins et de ne pas se limiter aux rapports imposés par les fournisseurs. Ainsi mesure-t-on non seulement l'efficience des ressources électroniques à un niveau fin, par exemple le coût d'un article récent ou d'une archive, mais

[47] Un tableau récapitulatif des principaux éléments de comparaison est présenté en annexe 14.

aussi l'efficacité ou encore la manière dont travaillent les différents utilisateurs, et donc la qualité du service rendu par l'établissement.

Il est cependant important de construire les mêmes rapports que ceux des fournisseurs pour permettre d'effectuer des comparaisons avec les leurs. Ce point est essentiel pour les négociations avec les éditeurs car les établissements peuvent ainsi construire leur modèle économique selon leurs propres critères et sont mieux à même d'argumenter. Grâce à cette solution, les données listées dans la recommandation du Comité de suivi, y compris celles concernant les catégories d'utilisateurs (pour ce dernier point, une partie des résultats dépend de l'organisation des établissements), seront obtenues.

Mais à ce jour, tant que tous les éditeurs ou plates-formes répertoriés dans les établissements n'ont pas été testés, la certitude d'avoir un outil complètement fiable et procurant toutes les données attendues n'est pas assurée. Il reste un travail important d'informaticien avant de pouvoir tirer des conclusions définitives.

Par ailleurs, si cette solution est retenue, il faudra prévoir la maintenance de l'outil selon deux aspects : l'ajout de nouveaux fournisseurs, les modifications en fonction de l'évolution des plates-formes. Il semblerait raisonnable d'organiser une maintenance mutualisée entre les établissements intéressés, car cela représente un travail important ; de plus, les mêmes éditeurs se retrouvent généralement dans plusieurs établissements. Cela rentabiliserait le travail.

Il est à noter que cette solution demande l'installation d'un proxy dans chaque SCD ou établissement concerné et l'obligation pour les utilisateurs de

configurer leur PC pour passer par ledit proxy, qu'ils soient sur le site de l'Université ou en mode nomade.

Les rapports COUNTER donnent globalement satisfaction à l'exception des libellés qui sont en anglais, ce qui est logique puisque c'est un standard national américain. Peut-être est-ce pour cette raison qu'il n'y a pas d'éditeurs français labellisés COUNTER, ceux-ci publiant des rapports en français ? Cependant, la documentation fournie étant souvent très succincte ou même absente, il n'est pas certain d'obtenir des données comparables entre plusieurs éditeurs, les logiciels des plates-formes étant différents. Les audits (Voir premier chapitre) pourront-ils corriger ces difficultés ? Il est avéré que même des fournisseurs compatibles COUNTER produisent des rapports parfois erronés. C'est ce qui semble ressortir aussi des premiers tests effectués durant cette mission et des comparaisons réalisées avec les résultats des fournisseurs, même si une plus grande antériorité serait nécessaire pour être affirmatif.

Les rapports COUNTER ont l'avantage de respecter un standard, au moins dans l'absolu. Si ce standard était international, cela permettrait une comparaison des statistiques entre pays, élément intéressant pour les établissements élaborant leur politique documentaire. Il est néanmoins nécessaire d'y joindre un outil de gestion des rapports afin d'uniformiser les résultats, faute de quoi un travail fastidieux de synthèse devra être fait manuellement. Un outil d'extraction automatique des données parait indispensable. La préparation des résultats est donc plus difficile que dans le cas de mesures locales qui sont gérées selon un processus unique.

Il est à noter que les rapports COUNTER ne répondent pas entièrement aux attentes du Comité de suivi, en particulier en ce qui concerne les catégories d’utilisateurs, point essentiel pour la politique documentaire des établissements.

Conclusion

En préambule, il faut rappeler que cette étude reste partielle, les tests n'ayant pu commencer suffisamment tôt, faute de moyens, et n'ayant pu porter sur tous les types de documents, dans tous les domaines de la connaissance, ni sur une période suffisamment longue. Par ailleurs, l'actualité du sujet lui confère beaucoup d'intérêt mais invite aussi à la prudence, certains aspects technologiques, décrits dans les chapitres précédents, pouvant subir des changements rapides.

La problématique des statistiques et indicateurs d'usages des ressources électroniques, leur importance dans l'élaboration des politiques documentaires au niveau des établissements comme du ministère sont indéniables et ont été exposées dans l'introduction. Au cours de cette année de travail, l'attente très forte des SCD pour disposer de statistiques s'est souvent exprimée.

Au terme de cette étude, aucune solution « COUNTER/locale » ne semble donner entière satisfaction (voir annexe 14). La recherche d'une standardisation internationale est un impératif et l'adhésion récente du consortium Couperin à COUNTER est une ouverture intéressante dans ce sens. On peut ainsi espérer obtenir une adéquation meilleure entre les rapports demandés aux fournisseurs et les attentes des établissements. Si plusieurs pays européens effectuent la même démarche, le Code pourrait devenir un standard international et non plus seulement national. Comparer ses statistiques et indicateurs à ceux de ses collègues européens et même mondiaux est aujourd'hui un aspect important dans la construction d'une politique documentaire.

Aujourd'hui, en fonction de l'avancée de la standardisation et du petit nombre d'éditeurs compatibles COUNTER[48], la solution « mesures locales » semble, même en l'état des tests, mieux répondre aux besoins des établissements. Cette solution est plus conforme aux souhaits du Comité de suivi, et construit des rapports faciles à utiliser. Elle permet aux établissements de mieux maîtriser les données statistiques et donc d'être à même de négocier plus facilement avec les fournisseurs, ceci même si elle demande des moyens matériels et humains non négligeables, encore que l'exploitation des données dans le cas des statistiques des fournisseurs en exige aussi.

Dans les deux cas, étant donné la charge de travail que représente la maintenance de l'outil et la redondance des éditeurs entre les établissements, une mutualisation des tâches entre les partenaires intéressés paraît

[48] Même si un petit nombre d'éditeurs font l'objet d'un très grand nombre d'interrogations, il ne faut pas négliger les domaines de la connaissance où on trouve d'autres éditeurs, des éditeurs français…

souhaitable ; c'est d'ailleurs cette solution qui a été retenue pour les tests.

Peut-être faut-il aller, comme le fait l'Université de Münster, vers un système mixte, sachant qu'alors les établissements doivent s'équiper d'un serveur mandataire ?

Dans ce cas, il serait d'ores et déjà nécessaire d'exiger dans les contrats négociés, la présence d'une clause demandant la fourniture de statistiques COUNTER.

Cette clause est également intéressante pour les établissements souhaitant n'effectuer aucune mesure locale.

Qu'il s'agisse de la solution locale ou de la solution mixte, il faut rappeler que la compatibilité proxy et Shibboleth n'existe pas aujourd'hui. Les établissements souhaitant obtenir localement des statistiques pourront permettre une utilisation nomade de leurs ressources mais en imposant le passage par un proxy et un contrôle de type CAS.

Cette étude fait le point à un moment donné, dans un processus en cours. Il serait intéressant que quelques établissements, représentant les différents domaines de la connaissance, puissent tester la solution locale en grandeur réelle. Cette opération devrait se faire sur une année complète au moins, sur tous les fournisseurs de l'établissement, afin de pouvoir vérifier ou infirmer l'intérêt de cette solution, et de la comparer avec la solution COUNTER. Ce choix nécessite la présence d'informaticien(s) pour compléter les scripts, la structure des compteurs…

Une telle expérimentation serait également profitable aux membres du

consortium Couperin amenés à travailler au sein du projet COUNTER.

Au terme de cette mission, si une suite est souhaitée, trois pistes de travail peuvent être recommandées :

- Continuer les tests en STM et les effectuer en droit/économie et SHS
- Participer activement à l'élaboration d'un standard international
- Travailler sur une solution serveur intermédiaire avec accès nomade.

Liste des annexes

Annexe 1

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

ministère éducation nationale enseignement supérieur recherche

Direction
de l'enseignement
supérieur

Service
des établissements

Sous-direction
des bibliothèques et
de la documentation

DES/SDBD/CJ/CL/
n°
Affaire suivie par
Claude Jolly
Téléphone
01 55 55 79 00
Fax
01 55 55 79 03
Mél.
claude.jolly
@education.gouv.fr

110 rue Grenelle
75007 Paris 07 SP

Paris, le

Le Directeur de l'enseignement supérieur

à

Madame Sabine BARRAL, Conservatrice générale des bibliothèques

S/C de Monsieur le Président de l'Université Claude-Bernard Lyon I

Objet : Mission relative aux indicateurs d'usages des ressources électroniques.

Le poids des acquisitions des ressources électroniques dans le budget de la documentation des universités ne cesse de s'accroître : ainsi en 2004, plus de 15 millions d'euros ont-ils été consacrés à l'acquisition de ressources électroniques, ce qui représentait une augmentation de 17 % par rapport à l'année précédente. Paradoxalement la mesure de l'usage de ces ressources en ligne est mal connue, comme en attestent les taux de réponse à l'enquête ESGBU menée par le Ministère de l'éducation nationale en 2004 (pour 2003) et en 2005 (pour 2004).

Eu égard à la charge budgétaire que représentent ces acquisitions et à la mise en œuvre de la LOLF, la définition d'indicateurs utilisables par les établissements d'enseignement supérieur et de recherche et leur tutelle s'avère indispensable, qu'il s'agisse d'indicateurs de gestion ou d'indicateurs de performance. Ces indicateurs serviront aussi pour la conduite des négociations avec les fournisseurs comme pour l'évaluation de la politique documentaire par les établissements.

Je vous confie, en accord avec le président de l'université Claude-Bernard Lyon 1, une mission d'étude visant en premier lieu à dresser l'état de l'art en matière d'indicateurs d'usage de la documentation électronique en milieu universitaire (étudiants et enseignants-chercheurs). Vous vous attacherez à la fois à la définition des items pertinents et à la capacité des établissements à renseigner les rubriques correspondantes. Vous étudierez notamment les documents de normalisation ou de standardisation mis au point au niveau international comme au niveau national, ainsi que les expériences et réalisations menées dans les établissements ou réseaux d'établissements pionniers et attestées dans la littérature professionnelle. Vous recenserez et évaluerez aussi, d'un point de vue généraliste, les dispositifs techniques existants ou pouvant être implémentés dans les établissements pour collecter et assurer l'exploitation des données destinées à construire ces indicateurs afin d'associer à la grille d'indicateurs proposés les modalités techniques requises.

Cette mission s'effectuera en liaison avec la Sous-direction des bibliothèques et de la documentation (Bureaux B1 et B2), l'association Couperin et les membres de la commission de normalisation CN 8 du CG 46 de l'AFNOR. Un comité de suivi, composé de professionnels de la documentation, sera constitué auquel vous rapporterez régulièrement l'avancement de votre étude.

Un rapport d'étape me sera adressé le 31 mars 2006, à la suite duquel le périmètre de la mission pourra être précisé.

Copie : M. Cavalier, directeur du SCD de l'Université Lyon I.

Annexe 2 : liste des sigles

Liste des sigles

AFNOR	Association française de normalisation
ARL	Association for research libraries
BIX	Bibliotheksindex
CCSD	Centre pour la communication scientifique directe
CNIL	Commission nationale de l'informatique et des libertés
COUNTER	Counting Online Usage of NeTworked Electronic Resources
CREPUQ	Conférence des Recteurs et des Principaux des Universités du Québec
CRI	Centre de ressources informatiques
ERE	Enquête sur les ressources électroniques
ERM	Electronic resource management
ERMS	Electronic resource management system
ESGBU	Enquête statistique générale auprès des services documentaires de l'enseignement supérieur
HAL	Hyper Article en Ligne
ICOLC	International coalition of library consortia
ICSTI	International Council for Scientific and Technical Information = Conseil International pour l'information scientifique et technique
IMAG	Institut de mathématiques appliquées de Grenoble
INIST	Institut de l'Information Scientifique et Technique
INRIA	Institut National de Recherche en Informatique et en Automatique
ISO	International standards organization
JUR	Journal use reports
LDAP	Lightweight Directory Access Protocol
LOLF	loi organique relative aux lois des finances
MINES	Measuring the impact of networked electronic services
NISO	National Information Standards Organization
SCD	Service commun de la documentation
SDBIS	Sous-direction des bibliothèques et de l'information scientifique
SHS	Sciences humaines et sociales
SICD	Service interétablissements de coopération documentaire
SIGB	Système de gestion de bibliothèque
STM	Sciences, Techniques, Médecine
Sushi	Standardized Usage Statistics Harvesting Initiative
TOC	Table of content
UCBL	Université Claude Bernard Lyon 1
UTC	Université de technologie de Compiègne
VPN	Virtual Private Network ou réseau privé virtuel (RPV)

Annexe 3 : enquête auprès des établissements

Indicateurs d'usages des ressources électroniques (« petite enquête »)

Pour m'aider dans la mission sur les indicateurs d'usage des ressources électroniques que m'a confiée le ministère, je vous serais reconnaissante de bien vouloir répondre aux questions ci-jointes dans les 15 jours. Compte tenu des délais mais ces résultats me serviront pour un rapport d'étape à fournir fin septembre, je préfère une réponse rapide même si elle n'est pas exhaustive.
Avec tous mes remerciements pour votre collaboration et en vous demandant de bien vouloir m'excuser pour cette charge supplémentaire.

Statistiques des fournisseurs

De quelles statistiques disposez-vous ? merci de compléter le tableau ci-joint en cochant en rouge les éditeurs pour lesquels vous disposez de statistiques et en ajoutant les éditeurs manquants (pour en savoir plus sur Counter, vous pouvez consulter http://www.projectcounter.org)

Exploitez-vous ces statistiques,
- si oui, êtes-vous satisfait(e)
- si non, pourquoi ? (merci de l'expliciter par fournisseur)

Statistiques locales

Si vous avez vos propres statistiques,
- comment les obtenez-vous ?
- quel(s) outil(s) utilisez-vous pour extraire les données ? merci de donner les caractéristiques techniques en précisant s'il s'agit de fichiers web ou non
- est-ce le SCD qui les exploite ou est-ce le CRI ?
- utilisez-vous un proxy ?
  - se situe-t-il au niveau de l'université ?
  - se situe-t-il uniquement au niveau de la documentation électronique ?
  - si vous en utilisez un,

quelles sont les caractéristiques techniques ?
- système d'exploitation
- existence d'une fonction cache
- logiciel de paramétrage

y a-t-il beaucoup de bases de données non accessibles via le proxy ? si oui, merci de les citer

Exploitation des données

Pour exploiter vos statistiques à différents niveaux, utilisez-vous

- Excel
- XML et ses dérivés
- Des logiciels spécifiques : scholarlystats…
- Un ERM (environmental resources management) qui puisse intégrer les statistiques des ressources électroniques, si oui lequel ?
- …

(plusieurs réponses sont possibles)

Annexe 4 : membres du comité de suivi

Membres du comité de suivi

JOLLY Claude puis MARIAN Michel	SDBIS, Président du Comité
COLAS Alain	SDBIS
DUCLOS-FAURE Danièle	SDBIS
BRU Gaëla	SDBIS
NICOLAS Marie-Dominique	SDBIS
BERNON Jean	Lyon 3
BERTRAND Annie	UTC
CARBONE Pierre	Paris 12, Président du CN8, AFNOR puis Bureau Couperin (sept 2006)
CAVALIER François	Lyon 1, Bureau Couperin jusqu'à sept 2006
COBOLET Guy	BIUM
ETIENNE Catherine	Bordeaux 1, Bureau Couperin depuis sept 2006
OKRET-MANVILLE Christine	Paris Dauphine
REIBEL BIEBER Iris	ULP

Annexe 5 : Sélection bibliographique

Sélection bibliographique

Accounting for the authentication and authorization infrastructure : pilot study / The Swiss education & research network. – Switch, 2006.

ARL supplementary statistics 2003-04. – http://www.arl.org/stats/suppindex.htm

Benchmarking and statistics : « Cheap, useful and fairly valid » / Eve Woodberry. – Conférence IFLA, 2006, n° 105

Les bibliothèques universitaires, *In* Rapport public 2005 de la Cour des comptes (2006)

Bit by bit / Peter Webster, *In* Libraryjournal.com, Jan. 2006. – http://www.libraryjournal.com/article/CA6298564.html

BIX- the library index : working paper. – http://www.bix-bibliotheksindex.de, 2005

Capture usage with E-metrics / J.C. Bertot, C.R. McClure, D.M. Davis & J. Ryan, *In* Library journal, May 1, 2004, p. 30-32

Collecte de statistiques d'utilisation des revues : EBSCO finalise avec succès l'intégration de SUSHI / GFII, 2006

Communiqué de presse du 18/1/2006 de Swets : [Sushi]. – http ://www.swets.fr

Comptes rendus du groupe de travail « Consultation et utilisation des ressources électroniques » organisé par la SDBD (2003/2004)

Construire des indicateurs et tableaux de bord / sous la dir. De P. Carbone. – Enssib ; Tec&doc, 2002

COUNTER : an overview / P.T. Shepherd. – Usage statistics training seminar, Oxford, 9 December 2005. – http://www.uksg.org/presentations6/shepherd.ppt

COUNTER : making statistics useful / P. Shepherd, *In* ICSTI Winter Public Conference, London, 2007. –
http://www.projectcounter.org/ICSTI_2_January_2007.ppt

COUNTER 2005 / P. T. Shepherd, *In* Learned publishing, 18(4), 2005, p. 287-293

COUNTER and the development of meaningful measures / P.T. Shepherd. – ICOLC 2005, Poznan. – www.pfsl.poznan.pl/icolc/1/counter-2005.ppt

COUNTER codes of practice. – http://www.projectcounter.org/code_practice.html

COUNTER News & activities. – http://www.projectcounter.org/news.html

COUNTER – Code de bonne pratique. – http ://www.inist.fr/article57.html

COUNTER Release 2 compliance realized by Elsevier, *In* News and releases : press releases. – http://www.info.sciencedirect.com/news/press_releases/archive/archive2006/counter.asp

Développer des indicateurs de performance pour décrire les services et ressources électroniques dans les bibliothèques de recherche américaines / J.M. Maffré de Lastens. – Mémoire d'étude ENSSIB, 2001

Du côté de chez Swets : Spécial nouveautés électroniques : [Sushi et Scholarlystats] / Swets Information Services. – Mél du 2/2/06 ; www.scholarlystats.com ; http ://www.iwr.co.uk/articles

E-Metrics : Next Steps for Measuring Electronic Resources / by Julia C. Blixrud. – *In* ARL Bimonthly Report, 230/231, October/December 2003. – http://www.arl.org/newsltr/230/emetrics.html

E-Metrics for library and information professionals / A.White and E.D. Kamal. – Neal-Schuman, 2006. – 249 p.

eJournal interface can influence usage statistics : implications for libraries, publishers and project COUNTER / Philip M. Davis, Jason S. Price, *In* Journal of the American Society for information science and technology, 57, n°9, 2006 ; http://arxiv.org/ftp/cs/papers/0602/0602060.pdf

Electronic article and journal usage statistics (EAJUS) : proposal for an industry-wide standard / J. Cowhig, *In* Learned publishing, 14(3), 2001, p. 233-236

Electronic Resource Management = Gestion des ressources électroniques : [plaquette] / Innovative interfaces. – www.iii.com

Electronic resource usage statistics : defining a complex problem / Caryn Anderson. – http://web.simmons.edu/~andersoc/erus/ERUSlandscape.doc, 2006

Les enjeux de la publicisation des sciences sur internet / Nathalie Pignard-Cheynel. – Thèse, Grenoble 3, 2004

Equinox : library performance measurement and quality management system, performances indicators for electronic library services / P. Brophy, R. Poll, [et al.]. – http://www.equinox.dcu.ie/reports/pilist.html, 2000

ERM statistics with Sushi. – Innovative interfaces

ERUS : annotated bibliography. –http://web.simmons.edu/~andersoc/erus/bibliography.html

ESGBU : résultats 2003 et 2004 et documents pour la collecte 2005

Evaluating the usage of library networked electronic resources / T. Plum *In* International developments in library assessment and opportunities for Greek libraries technological education institution Thessaloniki, 2005. – http://www.libqual.org/documents/admin/PlumEvaluating%20Networked%20Electronic%20ResourcesGreece050527.doc

L'évaluation et les indicateurs de la performance des activités info-documentaires / Eric Sutter. – Paris : ADBS, 2006. – 60 p.

Expertise de ressources pour l'édition de revues électroniques : usages / A. Mahé, *In* Guide pour les revues numériques, http ://revues.enssib.fr/titre/5usages/3quantitatives/2traitement.htm

Extrait des réponses au CCTP du portail documentaire du Sudoc au sujet des proxies. – Archimed, 2003 (confidentiel)

Guidelines for statistical measures of sage of web-based information resources / ICOLC. – 2001. – http://www.library.yale.edu/consortia/2001webstats.htm

Helping you buy : Electronic resource management systems / S. Meyer, *In* Computers in libraries, 2005, 25 (10), p. 19-24

ICOLC Autumn 2005 meeting, Poznan. – http://www.pfsl.poznan.pl/icolc/agenda.html

Impact et usage des services électroniques des bibliothèques universitaires / Mondane Marchand ; Ebsco. – ABF, 2006

Innovative Press release, November 23, 2005

Intégration de Sushi / Ebsco, *In* Archimag, juin 2006, n° 195

ISO/DIS 2789 Information et documentation – Statistiques internationales de bibliothèques. – Nouvelle édition. – Afnor, 2006

ISO 11620 Information et documentation – Indicateurs de performance des bibliothèques : projet de 2ème édition. – Afnor
Lettre[s] d'information d'Ebsco, 2006

Lettre de mission janvier 2006

Managing the electronic collection with cost per use data / B. Franklin, *In* IFLA Journal, vol. 31, n. 3, 2005

Manuel théorique pratique d'évaluation des bibliothèques et centres de documentation / T. Giappiconi. – Ed. du Cercle de la librairie, 2001

Measuring the use of electronic library services, *In* NISO Z39.7-2004, Information services and use…, appendix B

Mesures à une échelle nationale / R. Poll. – Conférence IFLA, 2006, n° 105

MINES for libraries : measuring the impact of networked electronic services. – www.arl.org/stats/newmeas/mines.html

NESLi2 analysis of usage statistics : summary report / A. Conyers, P. Dalton. – http://www.ebase.uce.ac.uk

Le New measures Initiative de l'American Library Association (ARL) / M. Kyrillidou. –Conférence au Crepuq, Montréal, 2005

NISO Standardized Usage Statistics Harvesting Initiative (SUSHI) :
http://www.niso.org/committees/SUSHI/SUSHI_comm.html

Performance indicators for the digital library / by Roswitha Poll, *In* Liber Quarterly, 2001, 11, 244-258

Planning and evaluating library networked services and resources / ed. By J.C. Bertot and D.M. Davis. – Libraries unltd, 2004. – 354 p.

Pratiques d'évaluation des bibliothèques / S.Gier-Jeanmougin, L. Gramondi, N. Liess, C.B. Sene. – Rapport de recherche ENSSIB, 2004

Projet de loi des finances 2006

Proquest information and learning : technical support : What are the COUNTER usage reports ?. – http://proquest.com/techsupport/answers/lad/lad_ans_201.shtml

The publishers' perspective / D. Sommer. – COUNTER-UKSG Usage statistics Training Seminar, 27th June 2006. – http://www.uksg.org/presentations270606/sommer.pdf

Recent Developments in Electronic Resource Management in Libraries / Rafal Kasprowski, *In* Asis&T Bulletin, August/September 2006. –http://www.asis.org/Bulletin/Aug-06/kasprowski.html

Release 2 of the Counter Code of practice for journals and databases is published, *In* Access, ISSN 0217-5673, June 2005. – http://www.aardvarknet.info/access/number53/othernews2.cfm?othernews=29

Scholarly journal usage : the results of deep log analysis / D. Nicholas, P. Huntington and A. Watkinson, *In* Journal of documentation, 61(2), 2005, p. 248-280

Scholarlystats : pour collecter, uniformiser et analyser vos statistiques d'utilisation / Swets information service, 2006

The Standardized usage statistics harvesting initiative (SUSHI) / A. Chandler, T. Jewell, *In* Serials, 19(1), 2006, p. 68-70

Standards-Libraries, data providers, and SUSHI : the standardized usage statistics harvesting initiative / by A. Chandler, T. Jewell, *In* Against the grain, ISSN 1043-2094, April 2006 : Special pre-print. – http:///www.against-the-grain.com

Statistiques d'utilisation des ressources électroniques : le projet Counter / Chérifa Boukacem, Joachim Schöpfel, *In* BBF, 2005, n°4, p. 62-66

Les statistiques d'utilisations [sic] des ressources électroniques en ligne de l'INIST : vers une application du Data mining ? : mémoire de maîtrise / S. Claudel. – Inist, 2003, 60 p. + annexes

Successful Web survey methodologies for measuring the impact of networked electronic serves (MINES for libraries)… / B. Franklin. – Conference IFLA, 2005, n° 157

SUSHI: The NISO Standardized Usage Statistics Harvesting Initiative / B. Chawner, October 28th 2006. – http://litablog.org/2006/10/28/sushi-the-niso-standardized-usage-statistics-harvesting-initiative/

SUSHI : The technology unveiled : NISO Webinar, 2006 / Ted Fons, Innovative Interfaces ; Oliver Pesch, EBSCO Information Services

Update on DLF Electronic Resource Management Initiative (ERMI), with Focus on XML Schema for e-Resource Licenses / A. Chandler, *In* ALA 2004 Annual Conference.- http://209.85.129.104/search?q=cache:O4Pr_xBpy84J:www.library.cornell.edu/elicensestudy/dlfdeliverables/alaannual2004/ERMI-ala-annual-2004final.ppt+&hl=fr&ct=clnk&cd=2&gl=fr

Usage data for electronic resources : a comparison between locally collected and vendor-provided statistics / J. Duy and L. Vaughan, *In* The Journal of academic librarianship, 29 (1), p. 16-22

L'usage de la documentation électronique au Service de la documentation Lyon 1 / O. Poncin ; sous la dir. De P. Carbone. – Mémoire d'étude ENSSIB, 2004

Usage statistics of online journals / P. Van Laarhoven. – Congrès Liber, 2005
The use and users of scholarly e-journals : a review of log analysis studies / H.R. Jamali, D. Nicholas and P. Huntington. – *In* Aslib proceedings, 57, (6), 2005, p. 554-571

Verde : e-resource management / Ex-Libris. – http ://www.exlibris.fr/verde.htm

What we can learn about virtual scholars from usage data obtained from deep log analysis / D. Nicholas, T. Dobrowski, P. Huntington, *In* Congrès ICOLC, 2005

White paper on electronic journal usage statistics / J. Luther. – 2nd ed. – CLIR, 2001. – http://www.clir.org/PUBS/reports/pub94/contents.html

Annexe 6 : Glossaire

Glossaire

Page rapport	Sujet	Définition	Référence
16	EQUINOX	**EQUINOX** is a project funded under the Telematics for Libraries Programme of the European Commission. This project addresses the need of all libraries to develop and use methods for measuring performance in the new networked, electronic environment, alongside traditional performance measurement, and to operate these methods within a framework of quality management The project has two main objectives. Firstly, EQUINOX aims to further develop existing international agreement on performance measures for libraries, by expanding these to include performance measures for the electronic library environment. The second aim is to develop and test an integrated quality management and performance measurement tool for library managers	http://equinox.dcu.ie/
17	E-Metrics	The ARL **E-Metrics** work was established in 2000 to develop standard definitions for measures that libraries could use to describe: (a) the e-resources they make accessible, (b) the use made of the resources, and (c) the level of library expenditures The E-Metrics project is an effort to explore the feasibility of defining and collecting data on the use and value of electronic resources. Although ARL has some experience in tracking expenditures for electronic resources through the *ARL Supplementary Statistics*, there is a widely held recognition that more work needs to take place in this area	http://www.arl.org/newsltr/230/emetrics.html
17	MINES	Measuring the Impact of Networked Electronic Services (**MINES**) is an online transaction-based survey that collects data on the purpose of use of electronic resources and the demographics of users. As libraries implement access to electronic resources through portals, collaborations, and consortium arrangements, the MINES for Libraries(tm) protocol offers a convenient way to collect information from users in an environment where they no longer need to physically enter the library in order to access resources	http://www.arl.org/stats/newmea mines.html
21	DigiQUAL LibQUAL	Under the auspices of the Association of Research Libraries, the **DigiQUAL** protocol is being developed to assess the services provided for the user communities of the National Science Digital Library (NSDL) program. The existing **LibQUAL**+ TM	http://www.digiqual.org/digiqual/index.cfm

		protocol, currently used by over 600 libraries, is being modified and re-purposed for the parameters of online educational digital libraries and other digital collections. The work is supported by ARL and by NSF grant number DUE-0121769	
26	Fichiers journaux	**Fichiers journaux (fichiers logs ou logs)** : le terme **log** est notamment employé en informatique pour désigner un historique d'événements et par extension le fichier contenant cet historique Un **log (**ou **fichier log**) se présente sous la forme d'un fichier texte classique, reprenant de façon chronologique, l'ensemble des événements qui ont affecté un système informatique et l'ensemble des actions qui ont résulté de ces événements. Ainsi, pour un serveur de type Web, le fichier log regroupe à la fois les demandes d'accès à chacun des fichiers du serveur : - date et heure précise de la tentative d'accès ; - adresse IP du client ayant réalisé cet accès ; - fichier cible ; - et éventuellement système d'exploitation et navigateur utilisé pour cet accès. Le fichier contient également la réponse fournie par le serveur à cette demande d'accès (si le fichier est trouvé, le poids de celui-ci... sinon, le type d'erreur rencontré).	http://fr.wikipedia.org/wiki/Log http://www.dicodunet.com/definitions/hebergement/fichier-log.htm
26	Serveur mandataire	Un **serveur mandataire** est un serveur qui a pour fonction de relayer différentes requêtes et d'entretenir un cache des réponses. Connu en anglais sous le terme de « *Proxy server* ».S'il relie des requêtes http, il s'agit d'un **proxy Web**	http://fr.wikipedia.org/wiki/Serveur_mandataire
26	Cache	La mémoire **cache** (ou tout type de cache) est une mémoire intermédiaire dans laquelle se trouvent stockées toutes les informations que le processeur central est le plus susceptible de demander	fr.wikipedia.org/wiki/Cache
27	Data mining	Le **Data Mining** est un processus d'extraction de connaissances valides et exploitables à partir de grands volumes de données. Il a vocation à être utilisé dans un environnement professionnel et se distingue de l'analyse de données et de la statistique par les points suivants : -les techniques utilisées vont au-delà des techniques classiquement utilisées en statistiques : le Data Mining se situe à la croisée des statistiques, de l'intelligence artificielle, des bases de données. - les connaissances extraites par le Data Mining ont vocation à être intégrées dans	http://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es#D.C3.A9finition_g.C3.A9n.C3.A9rale

		le schéma organisationnel de l'entreprise ou de l'entité considérée. Le **Data Mining** est un ensemble de méthodes et techniques qui permettent la prise de décisions, à travers la découverte, rapide et efficace, de schémas d'informations inconnus ou cachés à l'intérieur de grandes bases de données. Ce n'est ni un système d'interrogation de bases de données, ni un système de statistique et de visualisation.	http://dess-droit-internet.univ-paris1.fr/bibliotheque/article.php3?id_article=143
27	Web mining	**Web mining** is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are **Web usage mining**, **Web content mining** and **Web structure mining**	http://computing-dictionary.thefreedictionary.com/Web+mining
30	Squid	Un serveur **Squid** est un serveur proxy performant, il est capable d'utiliser les protocoles FTP, HTTP, Gopher, et HTTPS. Contrairement aux serveurs proxy classiques, un serveur **squid** gère toutes les requêtes en un seul processus d'E/S, non bloquant **Squid** est un **logiciel** qui part du principe, que dans un même lieu les gens vont souvent voir les mêmes choses sur Internet. Il est donc judicieux d'éviter de rechercher à l'autre bout de la planète une page 5 fois de suite. C'est pourquoi on utilise un mandataire (un intermédiaire), qui au nom du "client" web va faire la recherche	http://fr.wikipedia.org/wiki/Squid http://stargate.ac-nancy-metz.fr/linux/cache/presentation/presentation.htm#definition
30	Query string	Les variables d'environnement CGI sont des variables transmises à un programme CGI, par le serveur Web l'invoquant, lors de son exécution. Elles fournissent des informations sur la requête effectuée par le client, sur le serveur et également sur le client lui-même. Par exemple, lorsqu'on effectue une recherche sur un site qui fonctionne avec des *CGI*, le script récupérera les termes de la recherche avec la variable d'environnement **« QUERY_STRING** ». Query string Contient tout ce qui suit le « ? » dans l'URL envoyée par le client. Toutes les variables provenant d'un formulaire envoyé avec la méthode « GET » sera contenue dans le *QUERY_STRING*	http://fr.wikipedia.org/wiki/Variables_d%27environnement_CGI
30	Code retour	Les **codes retour** sont importants car ils représentent le statut de la transaction. Le code de réponse est constitué de trois chiffres, le premier indique la classe de statut et les suivants la nature exacte de l'erreur. Les codes 20x indiquent que l'opération s'est correctement effectuée. Le	http://www.iprelax.fr/http/http_art5.php

		plus courant est le code 200 (OK)	
30	Double clic	Un utilisateur demande la version HTML du même article à quatre reprises aux intervalles de temps suivants : Requête 1: 9:51:10 Requête 2: 9:51:19 Requête 3: 9:51:32 Requête 4: 9:51:41 Si on applique le filtre des **doubles-clics** à l'exemple ci-dessus, on obtient le résultat suivant : la comparaison des requêtes 1 et 2 permet d'éliminer la requête 1 et de conserver la requête 2; puis, la comparaison des requêtes 2 et 3 permet de conserver les deux requêtes, puisque plus de 10 secondes se sont écoulées entre ces deux requêtes ; puis la comparaison entre la requête 3 et la requête 4 permet d'éliminer la requête 3 et de conserver la requête 4, puisque moins de 10 secondes se sont écoulées entre ces deux requêtes. Ainsi, en appliquant le filtre des doubles-clics à l'exemple ci-dessus, on obtient l'enregistrement de deux requêtes réussies	http://www.inist.fr/IMG/pdf/COUNTER_-_Code_de_Bonnes_Pratiques_v2d.pdf
33	XML	**XML** (Extensible Markup Language ou langage de balisage extensible) est un standard du World Wide Web Consortium qui sert de base pour créer des langages balisés spécialisés; c'est un « méta langage ». Il est suffisamment général pour que les langages basés sur XML, appelés aussi dialectes XML, puissent être utilisés pour décrire toutes sortes de données et de textes.Il s'agit donc partiellement d'un format de données	fr.wikipedia.org/wiki/XML
33	PERL	**PERL** signifie Practical Extraction and Report Language. Que l'on pourrait (essayer de) traduire par « langage pratique d'extraction et d'édition » Autre définition : PERL signifie Practical Extraction	http://www.med.univ-rennes1.fr/~poulique/cours/perl/perl_html/introperl02.html webmaster.lycos.fr/glossary/P/
36	Shibboleth	The **Shibboleth** software implements the OASIS SAML v1.1 specification, providing a federated Single-SignOn and attribute exchange framework. Shibboleth also provides extended privacy functionality allowing the browser user and their home site to control the Attribute information being released to each Service Provider. Using Shibboleth-enabled access simplifies management of identity and access permissions for both Identity and Service Providers	

42	NESLi2	**NESLi2** is the UK's national initiative for the licensing of electronic journals on behalf of the higher and further education and research communities, 2003-2006. NESLi2 is a product of the JISC and underwritten by the Higher Education Funding Council for England on behalf of the Funding Bodies	http://www.nesli2.ac.uk/
45	ERM	The purpose of an **ERM** (Electronic resource management) is to manage the lifecycle of electronic products…	http://litablog.org/2006/10/28/sushi-the-niso-standardized-usage-statistics-harvesting-initiative/

Annexe 7 : Rapports Counter

Rapports « COUNTER »

Revues et Bases de données (R2)

Rapports obligatoires

Journal Report 1 - Number of Successful Full-Text Article Requests by Month and Journal
(nombre de requêtes réussies portant sur de articles en texte intégral par mois et par revue)

Journal Report 2 - Turnaways by Month and Journal
(nombre de refus de connexion par mois et par revue)

Database Report 1 - Total Searches and Sessions by Month and database
(nombre total des interrogations et des sessions par mois et par base de données)

Database Report 2 - Turnaways by Month and database
(nombre de refus de connexion par mois et par base de données)

Database Report 3 - Total Searches and Sessions by Month and Service
(nombre total des interrogations et des sessions par mois et par service)

Rapports facultatifs

Journal Report 1a[1]: Number of Successful Full-Text Article Requests from an Archive by Month and Journal
(nombre de requêtes réussies portant sur de articles en texte intégral provenant d'archives par mois et par revue)

Journal Report 3 - Number of Successful Item Requests and turnaways by Month, Journal and Page Type [TOC, abs, Ref, full (pdf, ps, html), turnaways (pdf, html)]
(nombre de requêtes réussies portant sur les items et de refus de connexion par mois, par revue, par type de page)

Journal Report 4 - Total Searches run by Month and Service
(nombre total des interrogations par mois et par service)

Livres et ouvrages de références (R1)

Rapports obligatoires

Book Report 1 - Number of Successful Title Requests by Month and Title
(nombre de requêtes réussies portant sur un titre par mois et par titre)

[1] Ce rapport a été très récemment ajouté, sans qu'une nouvelle version n'ait été créée

Book Report 2 - Number of Successful Section (chapter...) Requests by Month and Title
(*nombre de requêtes réussies portant sur les parties par mois et par titre)*

Book Report 3 - Turnaways by Month and Title
(nombre de refus de connexion par mois et par titre)

Book Report 4 - Turnaways by Month and Service
(nombre de refus de connexion par mois et par service)

Book Report 5 - Total Searches and Sessions by Month and Title
(*total des interrogations et sessions par mois et par titre)*

Book Report 6 - Total Searches and Sessions by Month and Service
(*total des interrogations et sessions par mois et par service)*

Annexe 8 : Contacts

Contacts

Le Président de l'UCBL
Les Membres de la SDBIS participant au comité de suivi
Maud Arnaud, Ex Libris
Sylvie Belzile, Directrice du Service des bibliothèques, Université de Sherbrooke, Québec
Jean Bernon, directeur du SCD de l'Université Lyon 3 et son équipe
John Carlo Bertot, membre du TC46/SC8, animateur du GT4
Annie Bertrand et David Lewis, Université de technologie de Compiègne
Chérifa Boukacem, maître de conférences, Université de Lille 3
Pierre Carbone, directeur du SCD de l'Université Paris 12
François Cavalier, directeur du SCD de l'Université de Lyon 1 et son équipe
François Charbonnier, stagiaire en Master 2 de Science de l'Information et des Bibliothèques au SCD de l'Université Lyon 1
Daniel Charnay, directeur du CCSD (Hal)
Pierre Chourreu, directeur du SCD de l'Université de Toulouse 3
Guy Cobolet, directeur de la BIUM
Onil Dupuis, chargé de recherche principal, CREPUQ
Catherine Etienne, directrice du SCD de l'Université de Bordeaux 1
Marianne Giloux et Stéphane Rey, ABES
Benjamin Girard, Swets
Florent Guilleux et Olivier Salaün, CRU (Shibboleth)
Jean-François Jal, Professeur à l'UCBL
Heike Klingebiel, Springer
Carole Letrouit, SCD de l'Université Paris 5
Jean-François Lutz, SCD de l'Université de Metz
Modane Marchand, Ebsco
Nathalie Marcerou-Ramel, Couperin (jusqu'en janvier 2007)
Didier Mascarelli et José E. Ramos Silva, Elsevier
Cécile Mazet, Innovative

Christine Okret-Manville, SCD de l'Université Paris Dauphine

Françoise Pellé, directrice du Centre international ISSN et son équipe

Roswitha Poll, présidente du TC46/SC8 – ISO, animatrice du GT2

Catherine Poupon-Czysz, responsable du Département Portails et Services d'information, INIST

Gilles Rech et Benoît Janiaud, CISR Université de Lyon 1/Insa Lyon

Iris Reibel Bieber, directrice du SCD de l'Université Louis Pasteur-Strasbourg et son équipe

Serge Rouveyrol, IMAG, Grenoble

Philippe Russell et Laurent Perillat, SICD 1 de l'Université Joseph Fourier et de l'Institut polytechnique de Grenoble

J J Schwartz et T Simoni, CRI Université de Lyon 1

P. Shepherd, directeur de projet COUNTER

Bruno Van Dooren, directeur du SCD de l'Université de Paris 1 (jusqu'en septembre 2006)

Annexe 9

Données statistiques et indicateurs
Tableau des items retenus par le Comité de suivi
et définitions

03/08/2007

	bases de données bibliogra-phiques	autres bases de données	périodiques électroniques et bases de données de périodiques en texte intégral payants	livres électroniques et autres documents numériques	documents numérisés produits localement	catalogue	ressources libres de l'internet	service de références bibliogra-phiques en ligne	site électronique de la bibliothèque
nombre de sessions par an	x	x	x	x		x	x		
nombre de sessions par mois	x	x	x	x	x	x	x		
nombre de sessions par an par catégorie d'usagers à desservir	x	x	x	x	x	x	x		
nombre de sessions par mois par catégorie d'usagers à desservir	x	x	x	x		x	x		
nombre de sessions par an et par titre	x	x	x	x	x	x	x		
nombre de sessions par mois et par titre	x	x	x	x		x	x		
nombre de sessions par an et par service	x	x	x	x	x	x	x		
nombre de sessions par mois et par service	x	x	x	x		x	x		
nombre de sessions rejetées par an	x	x	x	x	x	x			
nombre de sessions rejetées par mois	x	x	x	x	x	x			
nombre de sessions rejetées par an et par titre	x	x	x	x	x	x			
nombre de sessions rejetées par mois et par titre	x	x	x	x	x	x			
nombre de recherches (requête intellectuelle) par an	x	x	x	x		x	x		
nombre de recherches (requête intellectuelle) par mois	x	x	x	x	x	x	x	x	
nombre de recherches (requête intellectuelle) par an et par titre	x	x	x	x		x	x		
nombre de recherches (requête intellectuelle) par mois et par titre	x	x	x	x	x	x	x	x	
nombre de recherches (requête intellectuelle) par an et par service	x	x	x	x		x	x		
nombre de recherches (requête intellectuelle) par mois et par service	x	x	x	x	x	x	x	x	
nombre de recherches (requête intellectuelle) par an et par catégorie d'usagers à desservir	x	x	x	x		x	x		
nombre de recherches (requête intellectuelle) par mois et par catégorie d'usagers à desservir	x	x	x	x	x	x	x	x	
nombre d'unités de contenu documentaire téléchargées par an			x	x					
nombre d'unités de contenu documentaire téléchargées par mois			x	x	x				
nombre d'unités de contenu documentaire téléchargées par an, par titre et par date			x	x					
nombre d'unités de contenu documentaire téléchargées par mois, par titre et par date			x	x	x				
nombre d'unités de contenu documentaire téléchargées par an et par catégorie d'usagers à desservir			x	x					
nombre d'unités de contenu documentaire téléchargées par mois et par catégorie d'usagers à desservir			x	x	x				
nombre d'unités de contenu documentaire téléchargées par type de page** par an			x	x	x				
nombre d'unités de contenu documentaire téléchargées par type de page par mois			x	x	x				
nombre d'unités de contenu documentaire téléchargées par type de page par an et par titre			x	x	x				
nombre d'unités de contenu documentaire téléchargées par type de page par mois et par titre			x	x	x				

nombre d'unités de contenu documentaire téléchargées par type de page par an et par catégorie d'usagers à desservir			x	x	x				
nombre d'unités de contenu documentaire téléchargées par type de page par mois et par catégorie d'usagers à desservir			x	x	x				
nombre d'enregistrements téléchargés par an	x	x				x			
nombre d'enregistrements téléchargés par mois	x	x				x			
nombre d'enregistrements téléchargés par an et par titre	x	x				x			
nombre d'enregistrements téléchargés par mois et par titre	x	x				x			
nombre d'enregistrements téléchargés par an et par catégorie d'usagers à desservir	x	x				x			
nombre d'enregistrements téléchargés par mois et par catégorie d'usagers à desservir	x	x				x			
nombre de visites virtuelles par an									x
nombre de visites virtuelles par mois									x
nombre de visiteurs différents par mois						x			x
nombre de visites par origine géographique par mois						x			x
nombre de documents (collection électronique)	x	x	x	x	x				
nombre de documents par catégorie d'usagers à desservir	x	x	x	x	x				
satisfaction des usagers par catégorie d'usagers à desservir	x	x	x	x	x	x	x	x	x
pourcentage des titres demandés par rapport à la collection (papier et électronique)			x	x	x		x		
pourcentage des sessions rejetées	x	x	x	x		x			
coût par session	x	x	x	x					
coût par session pour chaque titre	x	x	x	x					
coût par requête	x	x	x	x					
coût par unité de contenu documentaire téléchargée		x	x	x					
coût par unité de contenu documentaire téléchargée pour chaque titre		x	x	x					
coût par catégorie d'usagers à desservir	x	x	x	x					
pourcentage des dépenses en fourniture d'information consacrées à la collection électronique	x	x	x	x					
coût par unité de contenu documentaire, par titre électronique et par éditeur	x	x	x	x					
taux d'utilisation des titres électroniques par éditeur (nombre d'unités de contenu documentaire téléchargées rapporté au nombre de titres par éditeur)	x	x	x	x					

données au niveau national
données au niveau établissement
données au niveau SCD
Couperin
objets gratuits
accès de l'extérieur

service=groupe de produits d'information en ligne protégé par une marque provenant d'un ou plusieurs fournisseurs, pour lequel on peut prendre un abonnement ou une licence et dont tout ou partie de la collection peut être interrogé

** résumés, requêtes TI pdf, HTML, PS…, TdM, bibl, refus de connexions TI pdf, html…

Annexe 10 **Directives COUNTER**

COUNTER Code of Practice

Journals and Databases
Release 2

Appendix D
Guidelines for Implementation

Introduction

For ease of reference, the numbering used in this Appendix corresponds exactly to that of the Code of Practice itself; where appropriate the relevant section of the Code of Practice text is quoted.

5a: *'Only successful and valid requests should be counted. For webserver-logs successful requests are those with a specific return code. The standards for return codes are defined and maintained by NCSA.'*

Requirement for Implementation:
Return codes that indicate a successful or valid request are specified in agreed, international web standards and protocols. The relevant governing document for hypertext protocols is RFC2068, which contains definitions for each Return Code number. There are five categories of return code numbers:

1xx (Information): this category provides information on a request, and often indicates that the user has come upon an experimental application.
2xx (Success): reserved for successful responses. This category of code is not usually seen by the user, but their browser will receive them and will know that whatever request was sent by the browser was received, understood and accepted.
3xx (Redirection): indicates the need for further action by the user's browser. User action may not be necessary, as the browser may deal with it automatically.

4xx (Client Error): this category of code is the one most frequently seen by the user and indicates an error.
5xx (Server Error): indicates where the server knows it has made an error, or is not capable of answering the request.

Categories **2xx** and **3xx** are relevant to Section 5a of the COUNTER Code of Practice, which deems that **only the following specific return codes indicate a successful or valid request:**

200 (OK) The request was successful and information was returned. This is, by far, the most common return code on the web.
304 (Not modified) In order to save bandwidth a browser may make a conditional request for resources. The conditional request contains an 'If-Modified-Since' field and if the resource has not changed since that date the server will simply return the 304 code and the browser will use its cached copy of the resource.

Requests that result in any other return codes within the 2xx and 3xx categories must not be counted. This exclusion covers:

206 (Partial content) This indicates that the server has only filled part of a specific type of request.
301 (Moved permanently): The addressed resource has moved, and all future requests for that resource should be made to the new URL. Transfer to the new location may be automatic or may require manual intervention by the user. Either way, a successful request to the new location will generate a 200 return code.
302 (Moved temporarily) This indicates that the content has moved while the page requested still has the same URL. The page is, therefore, not retrieved and must not be counted.
303 (See other) The response to the browser's request can be found elsewhere. Automatic redirection may take place to the new location.

Full information on the five categories of http return codes and their definitions may be found at: http://www.w3.org/Protocols/rfc2068/rfc2068 goto: Chapter 10 (pp 53-64): Status Code Definitions. More summarised information may be found at: http://www.cknow.com/faqs/What/404andOtherHTTPReturnCode.html .

5e. Guidelines for processing and filtering the raw usage data

The filtering of the 'raw' usage data needs to go through a number of consecutive steps in order to meet the COUNTER requirements.

Step1: Sorting the data file.

The file to be used for reporting should be sorted chronologically by user.
The following options for a user exist:

1. Where only the IP address of a user is logged that IP should be taken as the field to sort by.
2. When a session-cookie is implemented and logged, the session-cookie should be used to sort by.
3. When user-cookies are available and logged, the user-cookie should be used to sort by.
4. When the username of a registered user is logged, this username should be used to sort by.

Step 2: Remove all records with a return code other then '200' and '304'

Step 3: Run the 'double-click-removal' script

The following example illustrates how this script should work:

A user requests the HTML version of one and the same article four times within the following time intervals:

Request 1: 9:51:10

Request 2: 9:51:19

Request 3: 9:51:32

Request 4: 9:51:41

Applying the double-click filter to the above example has the following results: comparing Requests 1 and 2 removes Request 1 and retains Request 2; next, comparing Request 2 with Request 3, retains both Request 2 and Request 3 as more than 10 seconds have elapsed between these two requests; next, comparing Request 3 with Request 4 removes Request 3 and retains Request 4, as less than 10 seconds have elapsed between Requests 3 and 4. Thus, applying the double-click filter to the above example results in two Successful Requests being recorded.

COUNTER Code de bonnes pratiques

Revues et bases de données
Version 2
Annexe D
Directives pour la mise en œuvre

Trad. par N. Marcerou-Ramel
Septembre 2006

Introduction

Pour des raisons de commodité, la numérotation adoptée dans cette annexe reprend exactement celle du Code de bonnes pratiques ; lorsque que cela semblait opportun, la partie correspondante du Code de bonnes pratiques a été citée.

5a : "*Seules doivent être comptabilisées les requêtes réussies et valides. Pour les fichiers de connexion des serveurs Web, les requêtes réussies sont celles qui ont un code de retour spécifique. Les standards pour les codes de retour sont définis et tenus à jour par le NCSA.*"

Conditions de mise en œuvre :

Les codes de retour qui indiquent qu'une requête est réussie et valide sont spécifiés dans des normes et protocoles Web internationaux reconnus. Le document de référence pour les protocoles hypertextes est le RFC2068, qui contient les définitions de chacun des codes de retour. Il existe cinq catégories de numéros de codes de retour :

> **1xx** (information) : cette catégorie donne des informations sur une requête et indique souvent que l'utilisateur est tombé sur une application expérimentale.
> **2xx** (succès) : catégorie réservée aux requêtes réussies. Cette catégorie de codes n'est généralement pas vue par l'utilisateur, mais le navigateur de l'utilisateur reçoit ces codes et sait que la requête envoyée par le navigateur, quelle qu'elle soit, a été reçue, comprise et acceptée.

3xx (réorientation) : indique qu'une nouvelle action est requise de la part du navigateur de l'utilisateur. Une intervention de l'utilisateur n'est pas forcément nécessaire et le navigateur peut parfois traiter la tâche automatiquement.
4xx (erreur client) : cette catégorie de code est celle que l'utilisateur voit le plus fréquemment ; elle indique qu'il y a une erreur.
5xx (erreur serveur) : indique où le serveur sait qu'il a fait une erreur ou bien qu'il n'est pas capable de répondre à la requête.

Les catégories **2xx** et **3xx** s'appliquent à la Partie 5a du Code de bonnes pratiques COUNTER, selon laquelle seuls les codes de retour spécifiques suivants indiquent qu'une requête est valide ou réussie :

200 (OK) La requête est réussie et des informations ont été renvoyées. C'est de loin le code de retour le plus fréquemment rencontré sur le Web.
304 (Not modified ou non modifié) Afin d'économiser de la bande passante, un navigateur lance parfois une requête conditionnelle pour obtenir des ressources. La requête conditionnelle contient un champ 'If-Modified-Since' (« si modifié depuis ») et si la ressource n'a pas changé depuis la date indiquée, le serveur renverra simplement le code 304 et le navigateur utilisera sa copie de ressource disponible en cache.

Les requêtes qui donnent tout autre code de retour dans les catégories 2xx et 3xx ne doivent pas être comptabilisées. Cette exclusion s'applique notamment à :

206 (Partial content ou contenu partiel) Ce code indique que le serveur n'a répondu qu'en partie à un type spécifique de requête.
301 (Moved permanently ou déplacé de manière permanente) : la ressource recherchée a changé d'emplacement et toutes les requêtes ultérieures portant sur cette ressource doivent être adressées à la nouvelle URL. Le transfert vers la nouvelle localisation peut s'effectuer automatiquement ou nécessiter une intervention manuelle de l'utilisateur. Quoi qu'il en soit, une requête réussie dirigée vers la nouvelle adresse génèrera un code de retour 200.
302 (Moved temporarily ou déplacé de manière temporaire) Ce code indique que le contenu a changé d'emplacement alors que la page demandée a conservé la même URL. La page n'a donc pas été retrouvée et ne doit pas être comptabilisée.
303 (See other ou voir ailleurs) La réponse à la requête du navigateur peut être trouvée ailleurs. Il est possible d'être automatiquement redirigé vers la nouvelle adresse.

Des informations complètes concernant ces cinq catégories de codes de retour http et leur définition sont disponibles à l'adresse :
http://www.w3.org/Protocols/rfc2068/rfc2068 voir Chapitre 10 (pp 53-64) : Status Code Definitions. Des informations moins détaillées peuvent être consultées à : http://www.cknow.com/faqs/What/404andOtherHTTPReturnCode.html.

5e : directives pour le traitement et le filtrage des données d'utilisation brutes

Le filtrage des données d'utilisation « brutes » nécessite la mise en oeuvre d'un certain nombre d'étapes successives pour répondre aux spécifications de COUNTER.

Étape 1 : trier le fichier de données
Le fichier destiné au rapport doit être trié chronologiquement par utilisateur. Différentes options existent pour un utilisateur :

1. Lorsque seule l'adresse IP d'un utilisateur est connectée, alors c'est cette adresse IP qui est considérée comme la zone à utiliser pour le tri.
2. Lorsqu'un cookie de session est mis en œuvre et connecté, c'est le cookie de session qui doit être utilisé pour trier.
3. Lorsque des cookies utilisateurs sont disponibles et connectés, c'est le cookie utilisateur qui doit être utilisé pour trier.
4. lorsque le nom d'utilisateur d'un utilisateur enregistré est connecté, c'est ce nom d'utilisateur qui doit être utilisé pour trier.

Étape 2 : effacer tous les enregistrements avec un code de retour autre que « 200 » et « 304 ».

Étape 3 : lancer la procédure 'double-click-removal' ou « suppression du double-clic »

L'exemple suivant illustre comment cette procédure doit fonctionner :

Un utilisateur interroge la version HTML d'un seul et unique article quatre fois de suite dans les intervalles de temps suivants :

Requête 1 : 9 : 51 : 10

Requête 2 : 9 : 51 : 19

Requête 3 : 9 : 51 : 32

Requête 4 : 9 : 51 : 41

Après application du filtre sur le double clic, on obtient les résultats suivants : la comparaison des requêtes 1 et 2 conduit à supprimer la requête 1 et à retenir la requête 2 ; puis la comparaison de la requête 2 avec la requête 3 permet de retenir à la fois les requêtes 2 et 3 puisque plus de 10 secondes se sont écoulées entre ces deux requêtes ; puis la comparaison de la requête 3 avec la requête 4 permet de supprimer la requête 3 et de retenir la requête 4, étant donné que moins de 10 secondes se sont écoulées entre ces deux requêtes. Par conséquent, l'application du filtre sur le double clic à l'exemple ci-dessus a pour résultat l'enregistrement de deux requêtes réussies.

Recueil des données statistiques au SCD Lyon 1 Documentation technique

par François Charbonnier[1]

- I. Introduction

La construction d'indicateurs d'usage des ressources électroniques nécessite des mesures de son utilisation. Le recueil des données statistiques afférentes est effectué par l'outil « pAq » développé par M. Serge Rouveyrol de l'institut d'Informatique et de Mathématiques Appliquées de Grenoble et adapté et complété pour les besoins du Service Commun de la Documentation de l'Université de Lyon 1. Cet outil se situe encore à un stade expérimental de développement.

La mesure de l'utilisation de la documentation électronique peut s'appuyer sur un « code de bonnes pratiques » suivi par quelques éditeurs et en cours de standardisation, le code COUNTER. « pAq » a été développé pour recueillir des données conformes à cette recommandation afin de permettre des comparaisons avec les résultats des fournisseurs mais également pour répondre aux attentes du SCD de l'Université Lyon1.

L'utilisation de cet outil nécessite la connaissance des commandes de base UNIX, de l'écriture d'expressions régulières en langage PERL et également des connaissances en recherche documentaire et en statistique pour interpréter les résultats.

Ce document a été écrit pour le SCD de l'Université Lyon 1. Néanmoins, des remarques d'ordre général sur l'utilisation de « pAq » ont été intégrées sous la forme: *remarques.*
Les commandes ou lignes de commande seront mises en évidence ainsi : ***ligne de commande***

- II. Contexte général

- *1. Configuration du réseau informatique de l'université*

L'Université Claude Bernard Lyon 1 possède deux serveurs mandataires (ou proxies) fonctionnant avec un système d'exploitation GNU/Linux. Ils sont gérés par le Centre Inter-établissement pour les Services Réseaux. Ce sont des serveurs web qui servent de relais entre les réseaux locaux et internet. Le logiciel libre SQUID est utilisé pour le paramétrage des serveurs.

Une des fonctionnalités des serveurs mandataires est de conserver dans un fichier journal (ou fichier log) un historique de tous les événements survenus sur la machine. A chaque interrogation du proxy, une ligne indiquant notamment la date de la requête, sa provenance (adresse IP du poste de l'utilisateur) et le « statut » de la réponse (la « page » appelée s'affiche correctement, elle n'existe pas…), s'inscrit dans le fichier log.

[1] Stagiaire en Master 2 de Science de l'Information et des Bibliothèques au SCD de l'Université Lyon 1

L'accès aux ressources électroniques requiert une configuration spécifique des postes informatiques. Chaque utilisateur doit paramétrer son navigateur web à l'aide du fichier « proxy.pac ». Il suffit d'indiquer dans la fenêtre de paramètre de connexion l'adresse de configuration automatique du proxy : « http://www.univ-lyon1.fr/proxy.pac ».

Pourquoi cette configuration ?

Les éditeurs de ressources électroniques identifient le plus souvent aujourd'hui les utilisateurs qui ont accès à leurs documents par leur adresse IP. L'utilisation de proxies permet de ne déclarer auprès des éditeurs que leurs adresses. Aujourd'hui, peu d'éditeurs refusent cette configuration. Mais si c'est le cas, le proxy.pac permet d'accéder directement à l'éditeur via internet sans passer par le proxy. En effet, à chaque requête lancée par le navigateur, le proxy.pac, en fonction de l'URL, force le passage ou non par le proxy[2]. Il suffit donc d'indiquer dans ce fichier quelles adresses ne doivent pas passer par le proxy pour qu'elles ne soient pas redirigées.

Quels sont les intérêts de cette configuration ?

Ce dispositif permet d'imposer le passage par le proxy pour accéder à la documentation électronique. Ainsi, les fichiers logs vont contenir toutes les interrogations concernant les ressources numériques à l'exception de quelques éditeurs qui refusent cette configuration ou par choix de l'établissement.

Cette configuration est également intéressante au niveau de la gestion des adresses IP de l'université qui ont accès à la documentation électronique puisque seules les adresses des serveurs sont déclarées.

Un point sur le paramétrage des proxies :

Les serveurs de l'université doivent être paramétrés de façon spécifique pour la collecte des données. Ces paramétrages vont concerner trois points : la fonction cache, le « query string » des URLs et l'échelle de temps adoptée.

- Les proxies ont une fonction de « **cache** ». Ils gardent en mémoire les pages fréquemment visitées. Cette fonction doit être supprimée pour les URLs correspondant à la documentation électronique. En effet, il faut prendre en compte les interrogations satisfaites par les plates-formes des éditeurs et non par le cache. Ainsi, le travail comptabilisé du côté client, c'est-à-dire l'université, est le même que celui du côté fournisseur.

- Les mesures sur les données recueillies vont se faire grâce aux URLs et plus précisément au niveau du « **query string** ». Dans le cas d'une page web dynamique, c'est-à-dire générée par un programme, quand cette page est appelée, un programme construit la page à partir de données contenues dans l'URL. Cette partie de l'URL contenant ces dites données est appelée le « query string »[3]. Elle se repère par le point

[2] *Dans le cas de proxies dédiés à la documentation électronique, le proxy.pac est utilisé différemment*

[3] Une URL type est de la forme suivante : http://serveur/chemin/programme?query_string
Dans cet exemple : « http://www.rsc.org/ publishing/journals/JM/article.asp?doi=jm9960600573 » doi=jm9960600573 est le « query string ».

d'interrogation qui la précède. Il est donc nécessaire de conserver cette section de l'URL.

- Enfin, pour pouvoir être comparées, les statistiques du SCD et celles des fournisseurs doivent être calculées dans une période de temps identique. Il est donc important que l'échelle de temps adoptée pour le recueil des données soit la même que celle des fournisseurs. Il existe une convention. Toutes les dates doivent être en **UTC** ou « Temps Universel Coordonné » en Français. Aujourd'hui, ce point n'a pas encore été implanté à Lyon.

2. Constitution des fichiers journaux

Les fichiers journaux sont des fichiers textes dans lesquels s'inscrit une ligne à chaque requête sur le serveur. Une ligne d'un fichier log se présente ainsi :

1154419128.528 405 299.240.93.52 TCP_MISS/200 12220 GET http://www.rsc.org/publishing/journals/JM/article.asp?doi=jm9960600573 - DIRECT/135.116.140.15 text/html

Elle regroupe 9 éléments:

Éléments	Signification
1154419128.528	Date à laquelle s'est effectuée la requête. Elle est au format Epoch c'est-à-dire en secondes et calculé à partir du 1er janvier 1970.
405	Temps de réponse à la requête en milli-secondes.
299.240.93.52	Adresse IP « anonymisée » de l'utilisateur.
TCP_MISS/200	Cette colonne se compose de deux entrées séparées par un slash. La première entrée (TCP_MISS) indique sur quel port s'est faite la requête, et le statut de l'objet de la requête.[4]. La deuxième entrée (200) est un code retour qui indique le statut de la « page résultat »[5].
12220	Taille en octets du fichier délivré à l'utilisateur.
GET	Méthode pour passer les paramètres d'une requête http du navigateur au serveur.[6]
http://www.rsc.org/publishing/journals/JM/article.asp?doi=jm9960600573	URL demandée.

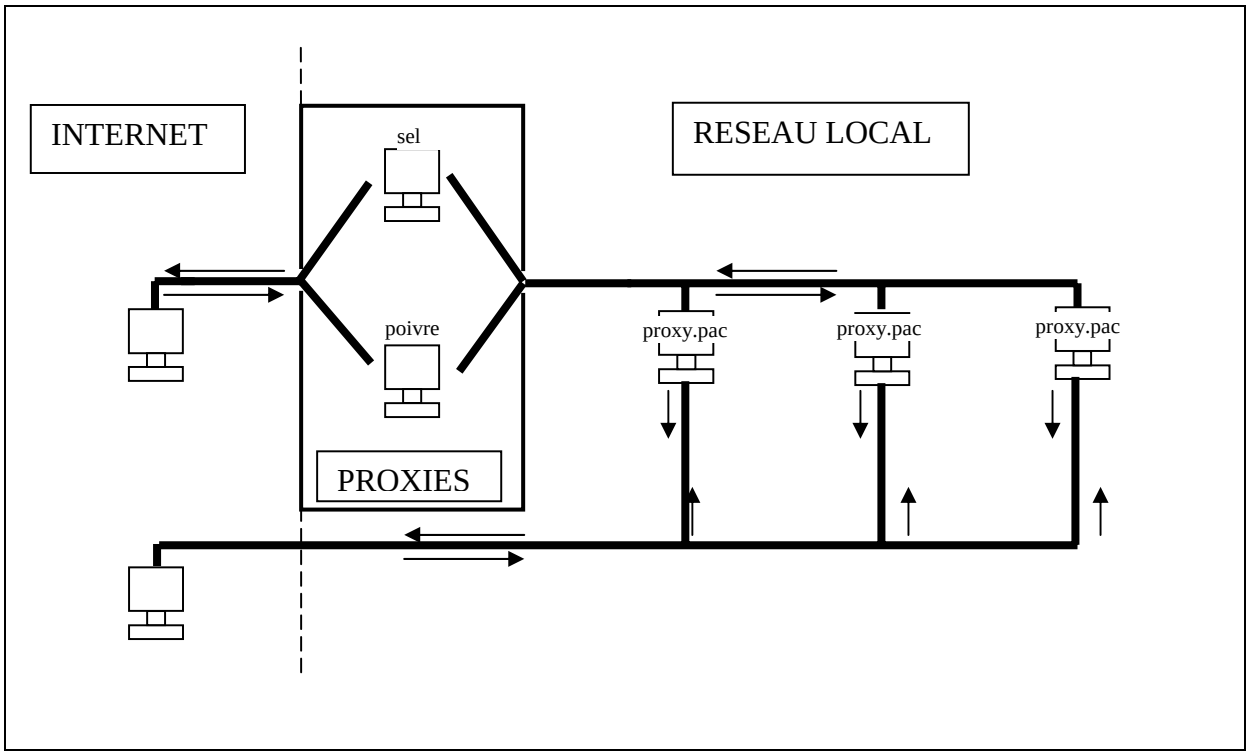
[4] Dans cet exemple, TCP_ signifie que la requête s'est faite sur le port « http », c'est-à-dire internet. On peut également trouver UDP_ correspondant au port ICP (Internet Cache Protocol). MISS signifie que l'objet de la requête n'était pas dans le « cache ». On peut trouver d'autres valeurs, comme HIT qui signifie que l'objet de la requête se trouve dans le « cache ». Pour plus d'informations, voir http://wiki.squid-cache.org/SquidFaq/SquidLogs.

[5] Ici, 200 signifie «affichage ok ». Exemples d'autres codes : 403 «affichage interdit », 404 « erreur sur la page »

[6] GET passe les paramètres directement dans l'URL alors que POST (autre méthode possible que l'on retrouve souvent dans les fichiers logs) passe les paramètres par une requête intermédiaire dont on ne retrouve pas la réponse dans les fichiers logs du proxy.

DIRECT/135.196.210.195	Cette colonne se compose de deux entrées séparées par un slash. La première entrée (DIRECT) indique la localisation de l'objet de la requête[7]. La deuxième entrée (135.196.210.195) n'est pas toujours renseignée. Elle correspond à l'adresse IP du serveur interrogé.
text/html	Type-mime : information sur le type de données contenues dans le résultat affiché[8].

3. Schéma récapitulatif[9]



- III. Récupération des fichiers journaux

La récupération des fichiers s'effectue quotidiennement, à partir d'un poste spécifique et de façon manuelle[10]. Pour les fichiers de la fin de semaine, il est possible de les récupérer uniquement le lundi qui suit. Il est important de penser à supprimer du serveur les fichiers récupérés pour ne pas surcharger l'espace disque du proxy.

---

7 Dans cet exemple, « DIRECT » signifie que l'objet a été cherché sur le serveur d'origine.

8 Exemples de types mimes : « application/pdf », « image/gif »…

9 *Cette configuration comportant deux serveurs mandataires est propre à Lyon1. Le système peut fonctionner également avec un seul serveur.*

10 Il est possible que cette récupération se fasse automatiquement. Cela impliquerait la mise en place par le CISR d'un script spécifique directement sur les serveurs.

Point sur la volumétrie (les données de Lyon1 sont prises comme exemple) :
Les deux fichiers logs compressés et concaténés font environ 400 Mo pour une journée de travail. En fin de semaine, la taille des fichiers est de l'ordre de 90 Mo.

Note :
Les commandes suivantes sont des commandes UNIX. Il est préférable pour les lancer sous Windows d'utiliser une invite de commande[11] comme Cygwin permettant de recréer un environnement Linux.

- *1. Connexion sécurisée aux serveurs de l'université :*

Quotidiennement, les fichiers logs sont copiés à 00:20 h par le CISR dans un répertoire spécifique des serveurs « sel » et « poivre »[12].

Pour accéder à ce répertoire, il faut se connecter en SSH[13] au serveur. Pour cela, le poste « récupérateur » a besoin d'un client SSH[14], c'est-à-dire un programme qui permet d'utiliser ce protocole. Du côté serveur, un compte doit être créé par le CISR sur lequel la connexion va être réalisée. Un login et un mot de passe seront alors attribués et permettront de se connecter avec la ligne de commande suivante :

ssh login@nom_du_serveur (sel ou poivre)

Une fois connectée, la commande « ***ls*** » permet de lister les différents fichiers présents dans le répertoire.
Pour fermer la session, il suffit de taper la commande « ***logout*** ».

La récupération des fichiers logs sur les serveurs « sel » et « poivre » s'effectue grâce à un transfert de fichier par SSH. Il s'effectue à l'aide d'une commande « scp » qui permet une copie sécurisée des fichiers du serveur jusqu'au poste client.

La ligne de commande est la suivante :
scp nom_du_dossier@serveur:nom_du_fichier /chemin_du_repertoire_de_destination

Une fois récupérés, les fichiers peuvent être supprimés du serveur. Il faut tout d'abord se connecter en SSH (***ssh login@nom_du_serveur***) puis utiliser la commande « ***rm nom_du_fichier***» et pour finir se déconnecter avec la commande « ***logout*** ».

---

[11] L'invite de commande est une fenêtre permettant d'exécuter des opérations en ligne de commande, c'est-à-dire une commande tapée au clavier.

[12] Les fichiers journaux sont constitués d'une journée à une autre de 00:20h à 00:20h. Or, ils doivent être constitués dans la tranche de temps utilisée par les fournisseurs, c'est à dire de minuit à minuit du jour suivant. Deux solutions sont possibles pour pallier à cela. Soit le CISR copie le fichier log à minuit dans le répertoire spécifique à la récupération. Soit, il faut reconstituer le fichier dans la tranche de temps correcte après la récupération. La première solution n'est pas réalisable, après renseignement auprès du CISR, l'heure de rotation des logs, c'est-à-dire la mise en place du nouveau fichier journal à la place du fichier de la journée qui se termine, a été mis en place après de nombreux tests. Son changement pourrait être problématique pour le fonctionnement du proxy. Ce point n'a pas été réglé à ce jour.

[13] SSH signifie « Secure Shell ». C'est un protocole qui permet une connexion sécurisée entre un serveur et un client SSH.

[14] Sous Windows, le logiciel libre « PuTTY » peut être utilisé, sous Linux, « openSSH ».

- *2. Concaténation des fichiers*

Les deux serveurs de l'université, « sel » et « poivre », fonctionnent en parallèle. Pour obtenir toutes les données d'une journée, il faut réunir les deux fichiers récupérés sur les deux serveurs. Les fichiers sont compressés. Pour les concaténer, il faut utiliser la commande :
zcat nom_du_fichier_sel nom_du_fichier_poivre | sort –n >nom_du_nouveau_fichier

La commande « ***zcat*** » permet de concaténer des fichiers compressés. Elle intègre la fonction de décompression, ainsi le fichier résultat est un fichier décompressé. Si les fichiers ne sont pas compressés, il faut utiliser la commande « ***cat*** » à la place de « ***zcat*** ».

La commande « ***sort –n*** » permet de trier chronologiquement les lignes du fichier résultat. Ce tri est nécessaire pour l'un des filtres qui sera ensuite appliqué au fichier.

Le fichier résultant est décompressé. Pour le compresser de nouveau, il faut utiliser la commande : ***gzip nom_du_fichier***

- IV. L'outil de traitement des données : « pAq »

L'outil de traitement des données créé par Grenoble s'appelle « pAq ». Il a été développé en langage PERL dans un environnement UNIX. Il peut être utilisé facilement sous Windows.

Note :
Les commandes décrites pour lancer les scripts sont des commandes UNIX. Il est préférable pour les lancer sous Windows d'utiliser une invite de commande comme Cygwin permettant de recréer un environnement Linux.

« pAq » se compose de cinq dossiers : 0.Readme, bin, conf, data, doc, lib et man.

- **0.Readme** comprend les fichiers « AUTHORS » et « COPYING » donnant pour le premier un historique des implications (personnes et travail effectué) dans le développement de « pAq » et pour le second, les détails du mode de distribution des logiciels sous GNU GPL (Licence Publique Générale).

- **bin** contient les scripts qui seront exécutés pour filtrer et analyser les données.

- **conf** se décompose en trois sous-dossiers : « id_TITRE_commun », « institutions » et « plateformes » et contient également deux fichiers de configuration, « filtre_periodiques.conf » qui sert au filtrage des données utiles au calcul des statistiques et « plateformes.conf » nécessaire au moment de l'analyse. Le dossier « plateformes » contient les sous-dossiers propres aux éditeurs. Chaque sous-dossier contient un fichier « package.pm »[15]. Il peut également contenir un fichier Excel, « id_TITRE » qui se compose de la liste des titres de l'éditeur et de la correspondance avec les codes (ISSN, coden ou autre) utilisés par ceux-ci pour les désigner. Le dossier « id_TTRE_commun » contient les fichiers « id_TITRES.xls » décrit ci-dessus

[15] L'utilisation de ces différents fichiers sera expliquée dans la partie V.

commun à plusieurs éditeurs. Actuellement, le fichier mis en place est celui des ISSN qui se localise dans le sous-répertoire « ISSN ». Le dossier « institutions » contient le fichier « ip_institutions.conf ». Il est utilisé dans l'analyse des catégories d'utilisateurs. Ce fichier permet de faire la correspondance entre les laboratoires et leurs adresses IP. Cette partie du travail fonctionne à Grenoble mais n'a pas été effectuée pour le moment à Lyon.

- **data** est un dossier qui permet de stocker des données.

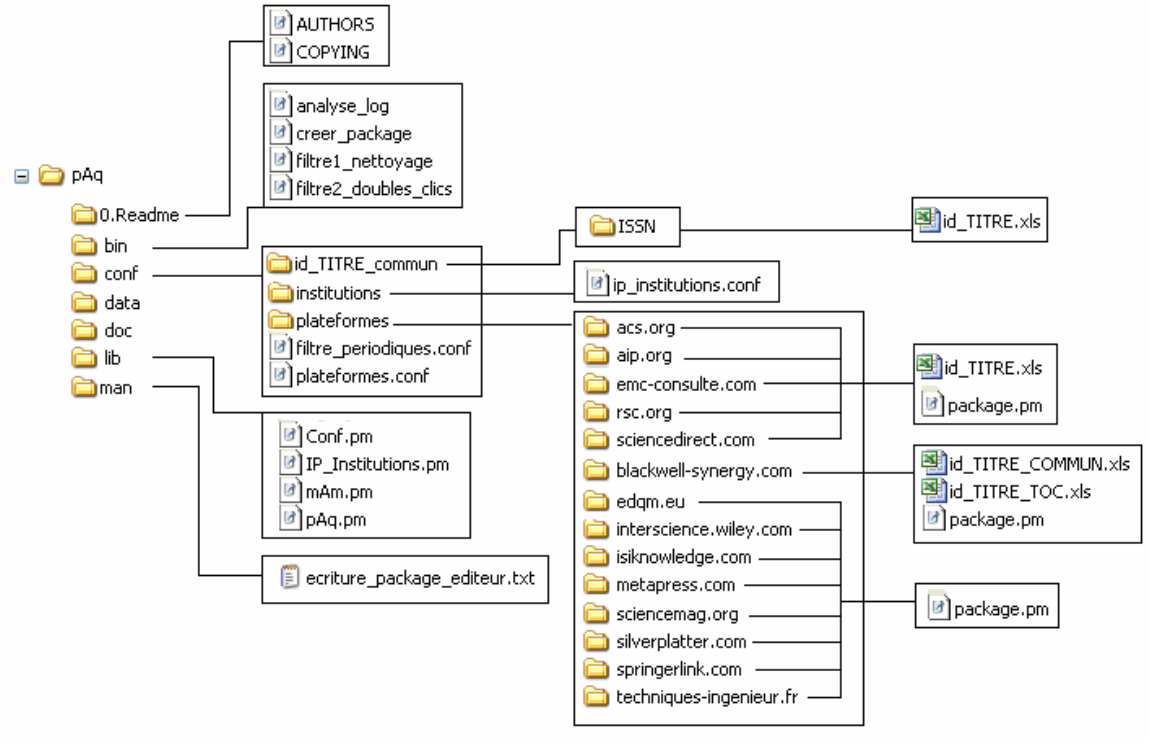
- **doc** est le dossier qui contient la documentation technique.

- **lib** contient les fichiers « pAq.pm », « Conf.pm ». Ces fichiers sont des modules composés de sous-programmes permettant l'analyse des logs. Deux autres modules sont également présents dans le dossier :« mAm.pm » et « IP_Institutions.pm ». Ils permettent avec le fichier « ip_institutions.conf » contenu dans le dossier « conf » de faire une analyse des catégories d'utilisateurs. Cette fonction n'ayant pas été étudiée jusqu'à présent à Lyon, l'utilisation de ces fichiers ne sera pas détaillée par la suite.

- **man** comprend une documentation sur l'écriture des fichiers « packages.pm » associés aux éditeurs. La commande suivante permet d'y accéder en ligne de commande : ***more $pAq/man/ecriture_package_editeur.txt***

Remarque :
Seuls les fichiers utilisés à Lyon et nécessaires pour la compréhension de l'outil seront décrits dans le présent document. Les dossiers de « pAq » peuvent contenir d'autres scripts. Pour connaître leurs usages, contactez M. Rouveyrol à l'origine du développement de « pAq ».

Vue du navigateur :



Installation :

L'installation de « pAq » sur un poste nécessite de modifier la valeur de la variable d'environnement. Cette manipulation est faite dans les fichiers suivants : « analyse_log » et « filtre1_nettoyage ». La variable d'environnement doit correspondre au chemin d'accès au dossier pAq.
La modification se fera dans la ligne suivante située dans les premières lignes du fichier :

```
$pAq = ($ENV{'pAq'}) ? $ENV{'pAq'} : "valeur à modifier";
```

V. Les scripts

- *1. Le pré-traitement*

Avant leur utilisation, les fichiers logs seront nettoyés. Seules les informations utiles aux mesures seront conservées. Ce nettoyage suit notamment les directives imposées aux fournisseurs déclarés conformes au code COUNTER mais également les besoins du SCD.
Les étapes de pré-traitement vont se décomposer en plusieurs parties.

Le script « filtre1_nettoyage » effectue les opérations suivantes :

- Éliminer du fichier toutes les requêtes adressées aux proxies ne correspondant pas à la documentation électronique. Pour ce faire, le filtre1 utilise un fichier de configuration « filtre_periodiques.conf ». Ce fichier contient les URLs des plates-formes des fournisseurs de documentation électronique. Ces URLs ont été tronquées de la manière suivante : en supprimant la partie précédant le premier « . » puis en éliminant celle suivant le nom du serveur.
  Par exemple, « http://www.rsc.org/Publishing/Journals/» deviendra « rsc.org ».
  Le filtre prend les URLs de chaque ligne du fichier log et les tronque de façon identique. Chaque URL est alors comparée aux lignes présentes dans les fichiers logs. Quand il y a une correspondance, la ligne est écrite dans le fichier résultat sinon, elle n'est pas prise en compte.

  Cette méthode évite d'avoir dans le fichier de configuration toutes les URLs des périodiques disponibles sur une même plate-forme (« http://pra.aps.org » ou « http://prb.aps.org » seront écrits « aps.org » par exemple). De plus, elle permet de filtrer « plus largement » et de prendre en compte certains changements d'URLs des plates-formes. Par exemple, fin 2006, les éditions « Techniques de l'ingénieur » ont modifié la configuration de leur plate-forme. Les documents ne sont plus uniquement accessibles via l'URL « http://www.techniques-ingenieur.fr » mais également par « http://pdf.techniques-ingenieur.fr »[16]. Le filtre permet de prendre en compte ces changements. Néanmoins, ce procédé est limité. Il ne permet pas de prendre en compte les URLs de la forme suivante « http://freemedicaljournals.com/ ». En appliquant le procédé de troncature utilisé par le filtre, la partie de l'URL à conserver est « com ». Le filtre perd alors de son efficacité.

[16] Souvent, on trouve plusieurs URLs pour le même éditeur qui différencie recherche et stockage des articles par exemple. Ainsi, pour l'éditeur techniques de l'ingénieur, la partie recherche s'effectue à l'adresse suivante : *http://www.techniques-ingenieur.fr* alors que les fichiers pdf sont téléchargés à partir de *http://pdf.techniques-ingenieur.fr*.

Il est possible d'effectuer une troncature « plus faible » qui ne supprime que la partie suivant le nom du domaine. L'avantage de cette troncature est de prendre en compte toutes les URLs mais elle nécessite au niveau du fichier de configuration du filtre de déclarer toutes les URLs des périodiques d'une même plate-forme par exemple (dans le cas où l'URL se construit ainsi : *http://periodique.serveur.com).* L'inconvénient de cette méthode est qu'il faut modifier le fichier de configuration quand les plates-formes changent d'URL mais également quand il y aura suppression d'un abonnement ou mise en place d'un nouvel accès. Il est donc nécessaire d'effectuer un suivi des plates-formes et de conserver les fichiers journaux bruts au moins 2 mois pour pouvoir comparer les résultats avec ceux du mois précédent.

En conclusion, tant que ce problème n'est pas résolu, il est nécessaire de garder les fichiers logs bruts afin de pouvoir les re-filtrer en cas de modification du script.

- Parmi les requêtes restantes, nous avons besoin de ne garder que celles faisant référence aux documents pdf, html, rtf et postscipt ainsi que les images aux formats tiff et jpeg (pour les documents numérisés en mode image). Pour repérer ces éléments, le filtre1 utilise les types-mimes, placés à la fin de chaque ligne d'un fichier log.

- Selon le Code COUNTER, on ne retiendra que les requêtes avec les codes retour 200 ou 304. Elles correspondent aux requêtes réussies et valides. Néanmoins, il pourra être également intéressant de garder à part certaines lignes avec un code retour différent de ceux conseillés par la recommandation COUNTER. Par exemple, pour mesurer les refus de connexion ou « turnaways »[17] sur la plate-forme de Springer, il est nécessaire de repérer les lignes ayant le code retour « 302 ».

```
1166433466.392      991      35.315.61.7      TCP_MISS/302      662      GET
http://www.springerlink.com/content/g5r826351421h401/resource-secured/
?target=fulltext.pdf - DIRECT/173.433.9.250 text/html
```

Le fichier « filtre2_doubles_clics » effectue l'opération suivante :

- Prendre en compte les doubles clics. Le code COUNTER propose d'éliminer les requêtes qui sont effectuées moins de 10 secondes après le premier clic dans le cas des pages html et moins de 30 secondes dans le cas des documents pdf.

> *« L'**exemple** suivant illustre comment appliquer ce script :*
>
> *Un utilisateur demande la version HTML du même article à quatre reprises aux intervalles de temps suivants :*
>
> *Requête 1: 9:51:10*
>
> *Requête 2: 9:51:19*
>
> *Requête 3: 9:51:32*
>
> *Requête 4: 9:51:41*
>
> *Si on applique le filtre des doubles-clics à l'exemple ci-dessus, on obtient le résultat suivant : la comparaison des requêtes 1 et 2 permet d'éliminer la requête 1 et de conserver la requête 2; puis, la comparaison des requêtes 2 et 3 permet de conserver les deux requêtes,*

[17] Les « turnaways full text» correspondent aux affichages indiquant qu'un document est inaccessible par manque de droits d'accès. Ils sont à distinguer des « turnaways » de session qui correspondent selon la définition du code COUNTER à « une tentative infructueuse de connexion à un service électronique, en raison du dépassement du nombre d'utilisateurs simultanés autorisé par la licence ».

puisque plus de 10 secondes se sont écoulées entre ces deux requêtes ; puis la comparaison entre la requête 3 et la requête 4 permet d'éliminer la requête 3 et de conserver la requête 4, puisque moins de 10 secondes se sont écoulées entre ces deux requêtes. Ainsi, en appliquant le filtre des doubles-clics à l'exemple ci-dessus, on obtient l'enregistrement de deux requêtes réussies. »[18]

Point sur la volumétrie (les données de Lyon1 sont prises comme exemple) :
Les fichiers bruts d'un mois entier ont une taille d'environ 7Go et 35Go une fois décompressée. La part documentaire de ces fichiers correspond à 460 Mo. Une fois filtrés, les fichiers ont une taille moyenne de 87 Mo.
Ces données ont été calculées sur une période de trois mois (décembre, janvier et février)

- *2. L'analyse*

Les fichiers journaux sont analysés par le script « analyse_log ». Son fonctionnement nécessite l'utilisation d'un fichier de configuration « plateformes.conf », les modules (ou packages), c'est-à-dire les fichiers pAq.pm et Conf.pm contenus dans le répertoire « lib » et ceux des éditeurs « package.pm » contenus dans le répertoire « plateformes ».

Le fichier de configuration « plateformes.conf » se compose de la même liste d'URLs tronquées que celle du fichier « filtres_periodiques.conf ». Les plates-formes qui ne sont pas étudiées sont placées en commentaires. Pour ce faire, le caractère « # » est placé en début de ligne[19] (par exemple : #acm.org). Les lignes commentées ne seront pas lues par le script d'analyse.

Les modules des éditeurs « package.pm » sont composés des compteurs associés aux mesures à faire. Des expressions régulières[20] sont contenues dans les compteurs et permettent de repérer les données à mesurer. Ces données sont présentes dans les URLs des requêtes inscrites dans les fichiers journaux. Le script d'analyse parcourt le fichier journal, à chaque fois qu'une URL satisfait à une expression régulière, le compteur correspondant est incrémenté d'une unité. On distingue deux types de compteurs, les compteurs par éditeur et les compteurs par titre. En ce qui concerne les mesures par titre, à chaque incrémentation d'un compteur « titre », le compteur « éditeur » correspondant est également incrémenté. Ainsi, les compteurs par titre permettent d'avoir également les résultats par éditeur. Les titres sont repérés par des sous-expressions régulières qui sont localisées dans les expressions régulières des compteurs. Ces sous-expressions permettent de repérer les codes utilisés par les éditeurs pour repérer les titres comme l'ISSN, le coden ou un code spécifique aux plates-formes, afin de traduire les titres en clair. Les tableaux permettant les correspondances, titre/ISSN par exemple, sont contenus dans les fichiers Excel « id_TITRE ». Quand cela est possible, le compteur titre permet également d'obtenir des résultats par date. Quand celle-ci est présente dans l'URL, il suffit d'utiliser une deuxième expression régulière la désignant pour qu'elle soit prise en compte dans les résultats par titre.

---

18 COUNTER Code de bonnes pratiques. Version 2. Annexe D. Directives pour la mise en œuvre - http://www.inist.fr/IMG/pdf/COUNTER_-_Code_de_Bonnes_Pratiques_v2d.pdf , 2005.

19 Caractère utilisé par le langage Perl pour écrire des commentaires dans un programme.

20 Les expressions régulières décrivent des schémas (c'est-à-dire des ensembles de chaînes de caractères).

Les modules pAq.pm et Conf.pm sont deux modules composés de plusieurs sous - programmes. Le module « pAq.pm » est le module « principal ». Il permet d'initialiser le programme d'analyse avec le fichier de configuration. pAq.pm utilise Conf.pm pour mettre en place les compteurs définis dans les « packages » des éditeurs. Tous les compteurs d'un module doivent être écrits dans le fichier Conf.pm. pAq.pm permet également l'incrémentation des compteurs et la création des fichiers de résultats « *pAq_editeurs.xls* » et « *pAq_titres.xls* ».

Le script d'analyse donne les résultats sous la forme de tableaux Excel : un tableau pour les résultats par éditeur, un autre pour les résultats par titre et date.

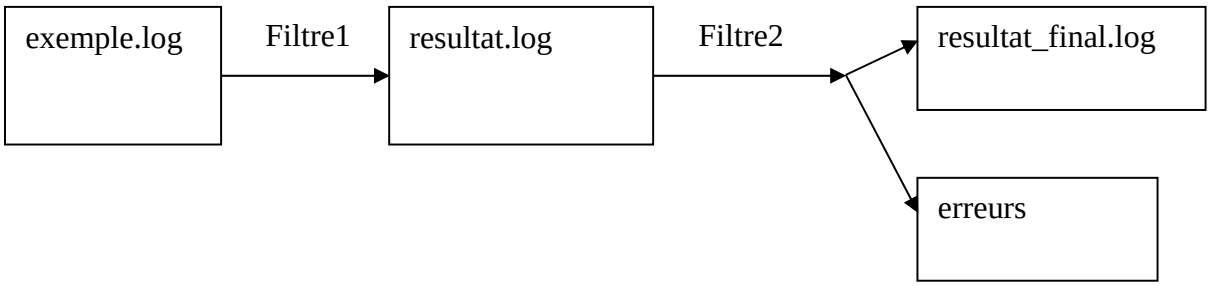
L'analyse doit être réalisée au moins tous les mois. Elle permet d'obtenir les résultats mensuellement et d'effectuer un suivi des plates-formes. En mesurant les variations entre les résultats des éditeurs et ceux que l'analyse donne, il est possible de déterminer si l'expression régulière d'un compteur est toujours juste. Une variation importante est le signal d'un changement de structure des URLs et probablement d'un changement de la configuration des plates-formes de l'éditeur.

- VI. Utilisation des scripts

- *1. Lancer les filtres*

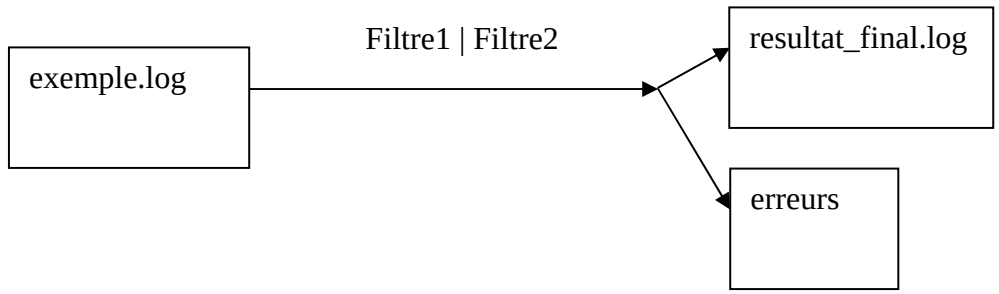
Dans une invite de commande, on peut soit lancer les deux scripts l'un après l'autre et utiliser le fichier de sortie du premier script (filtre1_nettoyage) comme fichier d'entrée du deuxième script (filtre2_doubles_clics), soit combiner les deux scripts à l'aide de l'opérateur « | ».

Dans le premier cas :



Les lignes de commandes sont les suivantes :

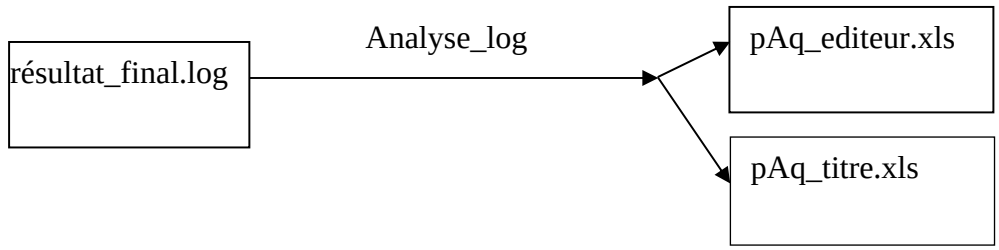
```
perl filtre1_nettoyage <exemple.log >resultat.log
perl filtre2_doubles_clics <resultat.log >resultat_final.log 2>erreurs
```



La ligne de commande est la suivante :
perl filtre1_nettoyage <exemple.log | perl2_doubles_clics >resultat_final.log 2>erreurs

- 2. Lancer l'analyse

Sur le résultat final, on peut lancer le script d'analyse :



La ligne de commande est la suivant e :
perl analyse_log <exemple.log

- 3. Création d'un module (package) pour un éditeur

La création d'un module pour un éditeur s'effectue en plusieurs étapes. Dans un premier temps, elle nécessite l'étude dans les fichiers logs des URLs des plates-formes des éditeurs. Le but est de repérer les chaînes de caractères caractéristiques des éléments à mesurer, c'est à dire de l'affichage des pages html présentant les résultats de recherches, des tables des matières, des résumés...

Exemple :
`http://pubs.acs.org/wls/journals/query/subscriber`**`Results`**`.html?op=searchJournals`
`http://pubs.acs.org/wls/journals/query/citationFind`**`Results`**`.html;jsessionid=`

Ces deux URLs correspondent à l'affichage des résultats positifs d'une recherche, ce qui correspond aux requêtes réussies selon le code COUNTER. On remarque que la chaîne de caractère « **Results** » se retrouve dans les deux lignes. Elle nous servira à écrire l'expression régulière contenue par le compteur « EDITEUR_SEARCH ».

Dans un deuxième temps, les résultats obtenus sont utilisés pour construire les expressions régulières contenues dans les compteurs.

Exemple :
En reprenant l'exemple précédent, on obtient le compteur suivant :
```
$regexp_EDITEUR_SEARCH = "^http://pubs[^\.]*\.acs\.org/.+Results";
```

- a. Décomposition des fichiers journaux :

- **En temps réel :**

La première étape du travail consiste à explorer la plate-forme d'un éditeur et à observer les résultats dans les fichiers journaux. Cette exploration nécessite de maîtriser l'outil de recherche documentaire afin d'envisager en premier lieu toutes les possibilités d'accès à un document. A chaque interrogation de la base de données, des lignes sont écrites dans les fichiers journaux. L'observation de ces lignes permet de comprendre quels sont les éléments qui font le lien entre les affichages des résultats d'une requête dans le navigateur et ce qui est inscrit au même moment dans le fichier log.

En pratique :
Pour que l'étude ne se fasse que sur un seul fichier log, il faut tout d'abord paramétrer son navigateur pour rediriger toutes les interrogations vers le même proxy[21]. Il suffit d'indiquer dans les « paramètres de connexion » de son navigateur l'adresse IP du proxy et le port utilisé pour la connexion.

Ensuite, il faut se connecter en SSH au serveur (voir III.1) et se positionner dans le répertoire contenant le fichier journal appelé « access.log ». L'adresse du répertoire est la suivante : « var/log/squid ».

La ligne de commande suivante permet d'afficher les 10000 dernières lignes inscrites dans le fichier « access.log ». La commande « grep » associée à la commande « tail » permet d'afficher uniquement les lignes contenant la chaîne de caractères passée en paramètre, par exemple une adresse IP:
```
tail -10000 access.log | grep XXX.XXX.XXX.XXX
```

Ainsi, une fois connecté au serveur, il est facile de lancer une requête sur la plate-forme de l'éditeur web et d'observer directement les lignes correspondantes s'inscrivant dans le fichier journal.

Remarque : Cette ligne de commande peut être associée à une commande « grep –v » permettant d'éliminer des résultats les lignes contenant une chaîne de caractères spécifique. Les images « gif » sont parfois très nombreuses et peuvent gêner la lecture. Le fichier log étant analysé en temps réel, il ne peut être filtré. Il est alors intéressant d'utiliser la ligne de commande suivante :

```
tail -10000 access.log | grep XXX.XXX.XXX.XXX | grep –v image/gif
```

[21] Les serveurs de Lyon1 fonctionnant en parallèle, la redirection par « proxy.pac » vers les proxies est aléatoire.

- **Sur une période d'un jour à un mois :**

Afin de compléter l'analyse, il est nécessaire de décomposer les fichiers logs sur une période allant d'une journée à un mois. Le but est de prendre en compte le maximum de cas de figures. Il s'agit par exemple de reconnaître un résultat de recherche, un article, une table des matières…

En pratique :

La commande « ***grep*** » permet de repérer des chaînes de caractères et de les afficher de manière spécifique. La commande « ***grep –v*** » permet d'éliminer les lignes contenant une chaîne de caractères et d'analyser ce qui reste.

Conseil : Les compteurs sont définis par type-mime « html » ou « pdf ». Extraire et placer dans un fichier les lignes de type-mime « text/html » et ainsi les séparer des lignes « application/pdf » permet de faire un premier tri sur les fichiers logs.

- b. Les compteurs et les expressions régulières : création du module :

Les compteurs déjà mis en place sont :

« Compteur éditeur »	« Compteur titre »
EDITEUR_SEARCH_HTML	TITRE_ABSTRACT_HTML
EDITEUR_SESSION_HTML	TITRE_ABSTRACT_PDF
EDITEUR_TURNAWAY_HTML	TITRE_TOC_HTML
EDITEUR_TELECH_HTML	TITRE_FULL_HTML
EDITEUR_ABSTRACT_HTML	TITRE_ARTICLE_HTML
EDITEUR_ABSTRACT_PDF	TITRE_SAMPLE_HTML
EDITEUR_TOC_HTML	TITRE_ASAP_HTML
EDITEUR_FULL_HTML	TITRE_FULL_PDF
EDITEUR_ARTICLE_HTML	TITRE_ARTICLE_PDF
EDITEUR_SAMPLE_HTML	TITRE_SAMPLE_PDF
EDITEUR_ASAP_HTML	TITRE_ASAP_PDF
EDITEUR_FULL_PDF	TITRE_ARCHIVE_PDF
EDITEUR_ARTICLE_PDF	TITRE_SUPPINFO_PDF
EDITEUR_SAMPLE_PDF	TITRE_FULL_PS
EDITEUR_ASAP_PDF	
EDITEUR_ARCHIVE_PDF	
EDITEUR_SUPPINFO_PDF	
EDITEUR_FULL_PS	

Il existe, en plus de ces compteurs, trois compteurs généraux « HTM », « PDF » et « PS ». Ils comptent respectivement, toutes les lignes se terminant soit par un type-mime « text/html », soit « application/pdf », soit « application/postscript ». Ils permettent de déterminer une limite maximum des résultats.

Dans le module « pAq.pm » qui réalise les calculs des compteurs, seuls les types-mimes « text/html », « application/pdf » et « application/postscript » ont été pris en compte. Ainsi, pour le moment, seul trois types de compteurs pourront être mis en place. Ils sont comptés

indépendamment les uns des autres. Le programme teste dans un premier temps les compteurs html pour toutes les lignes du fichier journal dont le type mime est « text/html » puis, les compteurs pdf pour celles dont le type-mime est « application/pdf » et enfin les compteurs ps pour les lignes avec le type-mime « application/postscript ».

Pour prendre en compte d'autres types-mimes, il faudra modifier « pAq » en conséquence.

Eléments à prendre en compte pour chaque compteur :

Les définitions présentes ci-dessous sont tirées du code de bonnes pratiques COUNTER pour les revues et les bases de données[22] à l'exception du compteur « TELECH » qui ne fait pas partie des rapports COUNTER. La définition donnée pour ce compteur est celle de la norme ISO 2789 « Statistiques internationales des bibliothèques ».

Pour chaque compteur, il ne faut prendre en compte que l'affichage des requêtes réussies.

Les compteurs « SEARCH » correspondent aux affichages des résultats positifs d'une recherche. Les pages des formulaires de recherche ainsi que les résultats nuls ne doivent pas être retenus. Par contre, un tri des résultats de recherche (par date, par auteur…) est compté comme un nouvel affichage.
Définition : «Une interrogation (search) est une requête intellectuelle spécifique, revenant classiquement à soumettre au serveur le formulaire d'interrogation du service en ligne. (EBSCO) »

Les compteurs « ABSTRACT » correspondent aux affichages des résumés.
Définition : «Un résumé est une courte présentation du contenu d'un article, incluant toujours ses conclusions. »

Les compteurs « TOC » correspondent aux affichages des tables des matières.
Définition : « *Revues* : une table des matières est une liste de tous les articles publiés dans le fascicule d'une revue. *Livres et ouvrages de référence* : une table des matière est une liste de tous les articles ou chapitres publiés dans le livre ou l'ouvrage de référence.»

Les compteurs « FULL » correspondent aux affichages des articles en texte intégral.
Définitions : «Un article en texte intégral est le texte complet, y compris l'ensemble de la bibliographie, des figures et tableaux d'un article, plus les liens vers tout autre document complémentaire accompagnant l'article.». Si les figures s'affichent indépendamment, elles ne sont pas comptabilisées dans les résultats.
FULL_HTML : «Article formaté en HTML pouvant être lu par un navigateur web. »
FULL_PDF : «Article formaté en PDF (portable document format) pouvant être lu avec le logiciel Acrobat Reader de Adobe ; tend à reproduire en ligne les pages d'un article telles qu'elles apparaissent dans la version imprimée. »
FULL_PS : «Article formaté en Postscript pour une reproduction fidèle à l'impression. »

Les compteurs « SESSION » correspondent au nombre de connexions à un service en ligne.
Définition : «Une session est une requête réussie sur un service en ligne. Il s'agit d'un cycle d'activité de l'utilisateur qui classiquement débute lorsque l'utilisateur se connecte au service ou à la base de données et qui se termine de façon explicite (en quittant le service par le menu

[22] Source : http://www.inist.fr/IMG/pdf/COUNTER_-_Code_de_Bonnes_Pratiques_v2a.pdf

quitter ou bien par une déconnexion), ou implicite (déconnexion automatique après une période de non utilisation). (NISO) »

Les compteurs « TURNAWAY » correspondent aux refus de connexion.
Définition : «Un refus (session rejetée) est défini comme une tentative infructueuse de connexion à un service électronique, en raison du dépassement du nombre d'utilisateurs simultanés autorisé par la licence. »

Les compteurs « TELECH » correspondent aux enregistrements de notices bibliographiques téléchargés.
Définitions : «Un *enregistrement descriptif* est un enregistrement bibliographique ou autre, traité en informatique dans un format normalisé, qui se réfère à et/ou décrit un document numérique ou une unité de contenu documentaire. Un ensemble d'enregistrements descriptifs est habituellement publié sous forme de base de données.»
« Un *téléchargement* est une demande aboutie d'un enregistrement descriptif ou d'une unité de contenu documentaire, par exemple pour affichage à l'écran, impression, sauvegarde ou envoi par message électronique ».

Compteurs spécifiques à l'éditeur ACS :

L'éditeur ACS distingue plusieurs types d'article. Les « articles » correspondent à la définition générale. Les « asap » correspondent aux articles récemment publiés sur internet. Les « sample » correspondent à des échantillons d'articles gratuits. Les « archives » comprennent tous les articles ACS de 1879 jusqu'à 1995 inclus. Les « supp_info » correspondent aux documents complémentaires accompagnant les articles.

Les compteurs «ARTICLE» correspondent aux affichages d'un article dont l'URL est du type :
`http://pubs.acs.org/cgi-bin/article.cgi/cmatex/2003/15/i19/html/cm034372t.html`

Les compteurs «SAMPLE» correspondent aux affichages d'un article dont l'URL est du type :
`http://pubs.acs.org/cgi-bin/sample.cgi/orlef7/2006/8/i05/abs/ol052861o.html`

Les compteurs «ASAP» correspondent aux affichages d'un article dont l'URL est du type :
`http://pubs.acs.org/cgi-bin/asap.cgi/nalefd/asap/html/nl060860z.html`

Les compteurs «ARCHIVE» correspondent aux affichages d'un article dont l'URL est du type :
`http://pubs.acs.org/cgi-bin/archive.cgi/ancham/1986/58/i02/pdf/ac00293a010.pdf`

Les compteurs «SUPP_INFO» correspondent aux affichages d'un article dont l'URL est du type :
`http://pubs.acs.org/subscribe/journals/jmcmar/suppinfo/jm060792t/jm060792tsi20060822_041833.pdf`

Remarques :
Les compteurs « ARTICLE », « SAMPLE » ou « ASAP » peuvent être soit des compteurs « PDF » soit des compteurs « HTML ». Les compteurs « ARCHIVE » et « SUPP_INFO » sont des compteurs « PDF ».

Pour un type donné, « PDF » ou « HTML », la somme de ces compteurs permet d'obtenir le compteur « FULL ».

Cas pratique pour le compteur SEARCH de ACS :
Les trois URLs suivantes correspondent à des résultats de recherche positifs. Les deux dernières correspondent à un tri des résultats de recherche qu'il faut compter comme un nouvel affichage.

```
http://pubs.acs.org/wls/preprints/preprintResults.html?op=searchPreprints
http://pubs.acs.org/wls/journals/query/query.html?op=refresh&sortSpec=date&
x=23&y=8&docsCount=50
http://pubs.acs.org/wls/journals/query/query.html?op=modifySearch
```

Les expressions régulières qui permettent de calculer le compteur SEARCH sont :

```
^http://pubs[^\.]*\.acs\.org/.+Results
^http://pubs[^\.]*\.acs\.org/.+op=refresh&sort
^http://pubs[^\.]*\.acs\.org/.+op=modifySearch
```

Remarque*:*

Les « TURNAWAYS FULL TEXT » décrivent les cas où l'utilisateur ne peut obtenir un élément, comme un article sous forme « pdf », par manque de droits d'accès.

Dans le cas d'un code retour 302:
Sur la plate-forme « springerlink », l'URL ci-dessous correspond au refus d'affichage d'un article sous forme « pdf »:

http://www.springerlink.com/content/g5r826351421h401/resource-secured/?target=fulltext.pdf

La chaîne de caractères « resource-secured » permet de caractériser cette URL. Cet exemple permet de montrer également qu'il faut faire attention à la lecture des URLs. Celle-ci se terminant par « .pdf » devrait correspondre à l'affichage d'un « pdf », pourtant, si l'on regarde la ligne du log la contenant, on s'aperçoit que le type-mime est « text/html ». Cette ligne correspond donc à l'affichage d'une page « html ». La page indique que l'utilisateur n'a pas accès à cette ressource.

```
1166433466.392    991 35.315.61.7 TCP_MISS/302 662 GET http://www.springer
link.com/content/g5r826351421h401/resource-secured/?target=fulltext.pdf -
DIRECT/173.433.9.250 text/html
```

La création du module :

Le fichier « creer_package » présent dans le répertoire « bin » contient la structure de base d'un module pour un éditeur. Il suffit de le placer dans un répertoire au nom de la plate-forme de l'éditeur, de le renommer « package.pm » et d'ajouter les compteurs à mettre en place pour qu'il soit opérationnel.

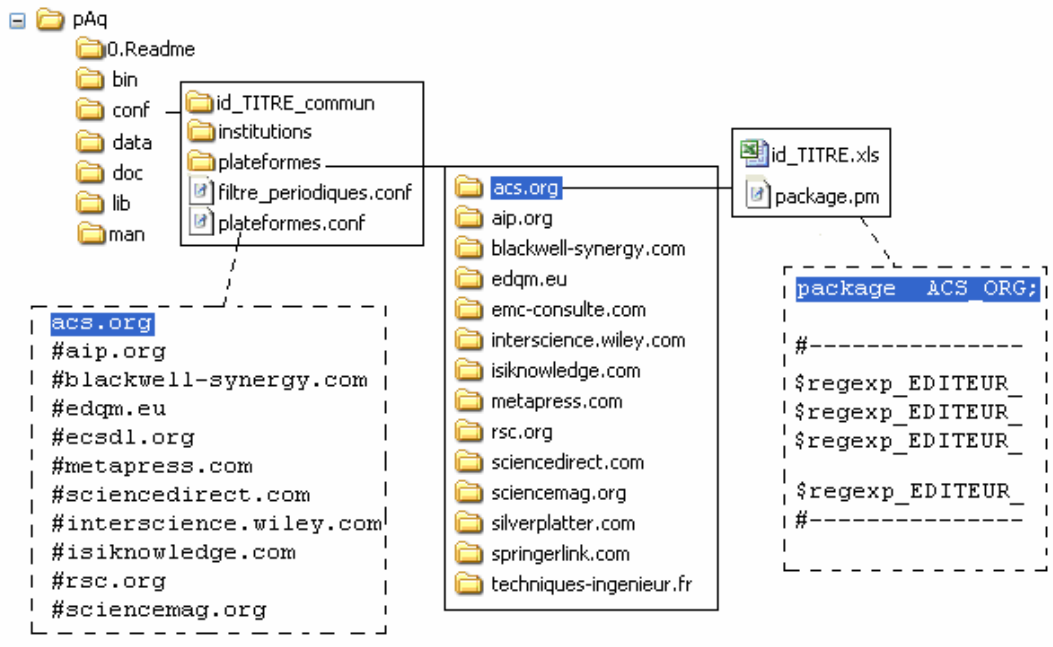
Les premières lignes du fichier sont les suivantes :

```
package  <REMPLACER CECI PAR LE NOM DU PACKAGE!>;

#-------------------------------------------------------------------------------
# METTRE EN PLACE LES COMPTEURS REALISABLES DANS LE MATCHING CHEZ UN EDITEUR
#-------------------------------------------------------------------------------

#-------------------------------------------------------------------------------
# format: <METTEZ ICI UN EXEMPLE CE CETTE URL>
# chercher puis ecrire la  regexp correspondante
#
# $regexp_EDITEUR_SEARCH_HTML = '';
#
```

Note sur le nom du répertoire contenant le module :

Le nom du répertoire doit correspondre à l'URL tronquée utilisée dans le fichier de configuration du filtre 1 et dans le fichier « plateforme.conf » qui indique quelles sont les plates-formes à étudier. Le nom du module est indiqué dans la première ligne du fichier. Cette ligne est du type « package ACS_ORG ». Le nom du module correspond au nom du répertoire dans lequel il se retrouve. Il est écrit en lettres capitales et les points « . » sont remplacés par le caractère souligné ou (underscore) « _ ».

Exemple : pour l'éditeur American Chemical Society :



Notes sur les expressions régulières :

Pour chaque expression régulière écrite, il est recommandé de garder en commentaire une URL ayant servi de modèle. Cela facilite le travail quand il faut reprendre les expressions pour affiner les résultats.
Les expressions régulières peuvent contenir des sous-expressions. Elles sont contenues dans des parenthèses. L'outil « pAq » a été conçu pour obtenir des résultats par éditeur mais également par titre. Ainsi, il sera possible quand une URL contient un code (ISSN, coden ou code propre à l'éditeur) désignant un titre, d'écrire une sous-expression régulière le décrivant. De plus, quand une date est également présente dans l'URL, une deuxième sous-expression pourra être écrite pour obtenir un résultat par titre et par date.

Exemple :
Pour l'URL suivante, l'expression régulière et ses sous-expressions seront :

```
http://pubs.acs.org/cgi-bin/article.cgi/cmatex/2003/15/i19/html/cm034372t.html

$regexp_TITRE_FULL_HTML =
'^http://pubs[^\.]*\.acs\.org/cgi-bin/article\.cgi/([a-z0-9]+)/([0-9]{4}] )
/.*/html/';
```

« [a-z1-9]+ » est la sous expression régulière décrivant le coden.
« [0-9]{4}]» est la sous expression régulière décrivant la date

Les sous-expressions régulières sont mises en mémoire dans une variable nommée $1 pour la première sous-expression, $2 pour la deuxième… Le plus souvent $1 correspond au titre et $2 à la date, ce qui signifie que le titre se trouve dans l'URL avant la date. Dans le cas contraire, si la date se retrouve avant le titre, il faut le signaler en attribuant la valeur « 1 » à la variable « $date_NOM_DU_COMPTEUR »

Exemple :
Avec l'URL suivante, il faudrait écrire :

```
http://pubs.acs.org/cgi-bin/article.cgi/2003/cmatex/15/i19/html/cm034372t.html

$regexp_TITRE_FULL_HTML =
'^http://pubs[^\.]*\.acs\.org/cgi-bin/article\.cgi/([0-9]{4}])/([a-z0-9]+)
/.*/html/';
$date_TITRE_FULL_HTML=1 ;
```

Les expressions régulières contenant les sous-expressions pour les titres et les dates seront contenues dans les « compteurs titre ». Les autres seront dans des « compteurs éditeur ».

Il y a 3 méthodes pour écrire une expression régulière[23]:

[23] Cette partie reprend les exemples du fichier «ecriture_package_editeur.txt » localisé dans le répertoire « man ».

1) La méthode « regexp_regexp » :
Elle est en général utilisée pour les compteurs de type éditeur. Les expressions régulières peuvent être écrites sous la forme suivante:

```
$regexp_EDITEUR_ABSTRACT = "^http://ieeexplore.ieee.org/xpls/abs_all.jsp";
```

S'il y a plusieurs expressions régulières (et **sans sous-expressions régulières**), on peut écrire :

```
$regexp1 = "^http://ieeexplore.ieee.org/xpls/abs_all.jsp";
$regexp2 = "^http://ieeexplore.ieee.org/abcd/abs_all.jsp";

$regexp_EDITEUR_ABSTRACT = "$regexp1|$regexp2";
```

2) méthode « regexp_array » :
S'il y a plusieurs expressions régulières (**avec des sous-expressions**), on peut écrire :

```
@liste_regexps = (
      "^http://ieeexplore.ieee.org/[^/]+/([0-9]+)/([0-9]{4})/",
      "^http://ieeexplore.ieee.org/[^/]+/abc/([0-9]+)/([0-9]{4})/"
)
$regexp_TITRE_FULL_PDF = \@liste_regexps;
```

Dans cette exemple $1 correspond au titre et $2 à l'année

3) méthode « regexp_fonction » :
La mise en place de certains compteurs nécessite l'utilisation de fonction, ce que l'on souhaite repérer ne pouvant être décrit par une expression régulière simple.

Dans l'exemple suivant, on souhaite compter les URLs de ce type :

```
http://www.sciencemag.org/cgi/content/short/313/5785/314
```

Et ne pas prendre en compte celles-ci :

```
http://www.sciencemag.org/cgi/content/full/291/5510/1965/F2
```

```
$regexp_TITRE_FULL_HTML=\&traiter_TITRE_FULL_HTML;

sub traiter_TITRE_FULL_HTML {
my $url = shift;
$regexp_TITRE_FULL_HTML  =
"^http://www\.sciencemag\.org/cgi/content/(full|short)";

 if (($url =~ /$regexp_TITRE_FULL_HTML/ ) && ($url !~ /\/[A-Z]+[0-9]+$/)) {
   my $id = $1;
   return (1,$id,0);
 }
 return (0,0,0);
}
```

Notes sur les compteurs titre et les tables de correspondance « code/titre du périodique »

Les modules contenant des « compteurs titre » ont une et parfois plusieurs tables de correspondance code (ISSN, coden, autres)/titre du périodique. Ces tables sont contenues dans des fichiers Excel. Ces fichiers ont une extension « xls » mais doivent avoir été enregistrés au format texte.

En règle général, ce fichier « id_TITRE.xls » est contenu au même niveau que le fichier « package.pm » c'est-à-dire dans le répertoire de l'éditeur. Par défaut, les compteurs utilisent le fichier « id_TITRE.xls » lié au compteur « TITRE_FULL_PDF ».

Si un compteur fait appel à une table différente, le fichier la contenant aura le nom suivant : « id_TITRE_nom_du_compteur ».

Exemple :
Pour un compteur « TOC », le fichier se nommera : « id_TITRE_TOC »

Un troisième cas est possible. Un compteur utilisant la table de correspondance « ISSN/titre du périodique » nécessite de mettre en place un fichier spécifique. Le fichier « package.pm » contenant ce compteur sera accompagné d'un fichier « id_TITRE_COMMUN.xls ». Ce fichier indique, pour chaque compteur l'utilisant, le nom du dossier contenant cette table commune à tous les compteurs.

Exemple :
La table de correspondance « ISSN/titre du périodique » est localisée dans le répertoire : « pAq/conf/id_TITRE_commun/ISSN ».

Pour l'éditeur Blackwell, trois compteurs font appel à cette table :
- TITRE_ABSTRACT_HTML
- TITRE_FULL_HTML
- TITRE_FULL_PDF

Le répertoire « blackwell-synergy.com » va donc contenir le fichier « package.pm » avec ses compteurs et le fichier « id_TITRE_COMMUN.xls » contenant les lignes suivantes :

	A	B
1	TITRE_ABSTRACT_HTML	ISSN
2	TITRE_FULL_PDF	ISSN
3	TITRE_FULL_HTML	ISSN
4		
5		

L'ajout d'un compteur :

L'intitulé d'un compteur doit contenir trois informations, l'indication « titre » ou « éditeur » (TITRE_ ou EDITEUR_), son nom précis (ABSTRACT_, FULL_...) et sur quels types d'affichage le compteur sera calculé (HTML, PDF ou PS)

Exemples : voir VI.3.b « tableau des compteurs déjà mis en place ».

L'ajout d'un compteur nécessite la modification du module Conf.pm dans le répertoire « lib ». Au niveau de la table de hachage « `%type_compteurs` », il faut ajouter les lignes correspondant au nouveau compteur en respectant la syntaxe établie. Si le nouveau compteur est un compteur par titre, il faut également ajouter le compteur par éditeur. Quand le compteur titre est incrémenté, le compteur éditeur correspondant l'est également automatiquement.

Il faut également modifier le sous-programme « print_compteur » dans le module pAq.pm pour que le compteur ajouté soit présent dans le fichier résultat. Cette modification touche une partie importante du module pAq.pm. Il est nécessaire d'effectuer une copie de sauvegarde du module avant de le modifier. Si le compteur à ajouter est un compteur « éditeur », seule la partie « résultats par éditeur » sera à modifier. Sinon, les deux parties, « résultats par éditeur » et « résultats par titre » seront à modifier. Pour connaître les différentes parties du programme à modifier, voir l'exemple qui suit.

Exemple :
L'ajout du compteur « TITRE_TURNAWAY » nécessite de modifier :

Dans le module « Conf.pm » :

```
%type_compteurs =  qw (
EDITEUR_SEARCH_HTML     1
EDITEUR_SESSION_HTML    1
EDITEUR_TURNAWAY_HTML   1
[…]
EDITEUR_FULL_HTML       1
TITRE_ABSTRACT_PDF      1
TITRE_TOC_HTML          1
TITRE_TURNAWAY_HTML     1
TITRE_FULL_HTML         1
TITRE_ARTICLE_HTML      1
[…]
TITRE_FULL_PS           1
);

@compteurs = qw (SEARCH SESSION TURNAWAY FULL ABSTRACT TOC ARTICLE […]);
```

Dans le sous programme « print compteur » du module « pAq.pm » :

```
sub print_compteurs {
my ($r_res, $r_institution) = @_;

mkdir "pAq_resultats";
chdir "pAq_resultats";

# resultats par editeur

open (EDITEURS, "> pAq_editeurs.xls") ||
die "fichier pAq_editeurs.xls increable : $!";

  print EDITEURS "TOT_PS\t".
                 "TOT_HTM\t".
                 "TOT_PDF\t"    .
                 "E_SESSION_HTML\t".
                 "E_TURNAWAY_HTML\t".
                 "E_SEARCH_HTM\t".
[…]
                 "E_TELECH_HTML\t".
                 "PLATEFORMES\n";


for my $plateforme (keys %{$r_res}) {
   my $r =  $r_res -> {$plateforme};

    print EDITEURS $r -> {'TOTAL_PS'}                . "\t".
```

```
                    $r -> {'TOTAL_HTML'}               . "\t".
                    $r -> {'TOTAL_PDF'}                . "\t".
                    $r -> {'EDITEUR_SESSION_HTML'}     . "\t ».
                    $r -> {'EDITEUR_TURNAWAY_HTML'}    . "\t".
                    $r -> {'EDITEUR_SEARCH_HTML'}      . "\t".

[…]
                    $r -> {'EDITEUR_TELECH_HTML'}      . "\t".
                    $plateforme                       . "\n";
  }
close EDITEURS;

# resultats par titre

my %t_res;

for my $plateforme (keys %$r_res) {
    my $r = $r_res -> {$plateforme};

    for my $type_cpt_TITRE (keys %Conf::type_compteurs) {
      next if $type_cpt_TITRE =~ /^EDITEUR/;

      if ($r -> {$type_cpt_TITRE}) {
        for my $titre ( keys %{$r -> {$type_cpt_TITRE}}) {

          for my $annee (keys %{$r -> {$type_cpt_TITRE} -> {$titre}}) {

            my $tit = "$annee\t$titre";

            if (!$t_res{$tit}) {
              my %compteurs;
              $t_res{$tit} = \%compteurs;

              $t_res{$tit} -> {'EDITEUR'} = $plateforme;
            }

            $t_res{$tit} -> {$type_cpt_TITRE} =
                      $r -> {$type_cpt_TITRE}  -> {$titre} -> {$annee};
          }
        }
      }
    }
  }

open (TITRES, "> pAq_titres.xls") ||
die "fichier pAq_titres.xls increable : $!";

  print TITRES "T_TURN_HTM\t" .
               "T_ABST_HTM\t" .
               "T_ABST_PDF\t" .
               "T_TOC_HTM\t" .
               "T_FULL_PS\t" .
[…]
               "T_ARCHIVE_PDF\t" .
               "T_SUPPINFO_PDF\t" .
               "EDITEUR\t" .
               "ANNEES\t" .
               "TITRES\n" ;

  for my $titre ( sort keys %t_res) {
```

```
          $editeur = $t_res{$titre} -> {'EDITEUR'};

    print TITRES $t_res{$titre} -> {'TITRE_TURNAWAY_HTML'} . "\t".
                 $t_res{$titre} -> {'TITRE_ABSTRACT_HTML'} . "\t" .
                 $t_res{$titre} -> {'TITRE_ABSTRACT_PDF'   . "\t" .
[…]
                 $t_res{$titre} -> {'TITRE_ASAP_PDF'}      . "\t" .
                 $t_res{$titre} -> {'TITRE_ARCHIVE_PDF'}   . "\t" .
                 $t_res{$titre} -> {'TITRE_SUPPINFO_PDF'}  . "\t" .
                 $t_res{$titre} -> {'EDITEUR'}             . "\t" .
                 $titre                                    . "\n" ;
  }

close  TITRES;

print "Les fichier pAq_... ont ete crees dans le directory courant\n";
}
```

Mise en application (extrait du rapport de stage de F. Charbonnier)

Pour des raisons de confidentialité, les données présentées dans les exemples suivants ont été modifiées tout en gardant les proportions entre elles.

La valeur du coefficient de corrélation a été calculée à l'aide du tableur Microsoft Excel. Une fonction « COEFFICIENT.CORRELATION » la renvoie automatiquement.

Nous allons présenter trois situations d'analyse à partir de résultats sur les revues.
La première correspond à un coefficient de corrélation significatif et à une régression linéaire concluant à une adéquation des données locales et de celles de l'éditeur. Dans le deuxième cas, le coefficient de corrélation est significatif et la régression linéaire montre des divergences entre les données. Le dernier cas présente un coefficient de corrélation non significatif ce qui ne permet pas de comparer les données locales et celles des fournisseurs.

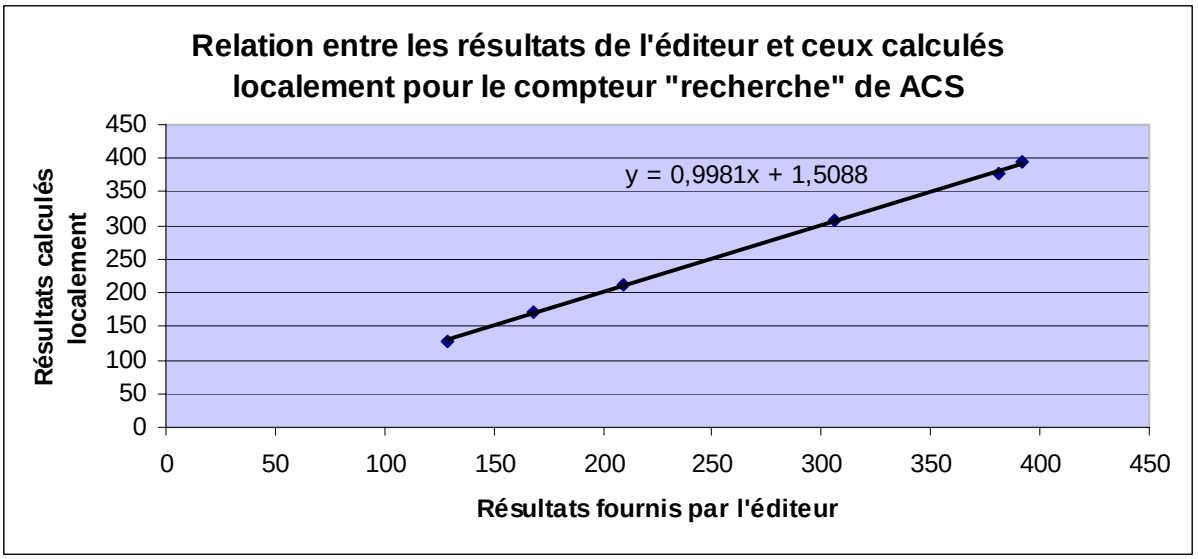
- *Première situation :*

Exemple : Compteur « recherche » de l'éditeur American Chemical Society

Résultats :

ACS	Recherche	
Mois	Résultats fournis par l'éditeur	Résultats calculés localement
Août 2006	129	128
Septembre 2006	210	211
Octobre 2006	392	396
Novembre 2006	382	379
Décembre 2006	168	171
Janvier 2007	306	308

Valeur du **coefficient de corrélation** : r = 0,9998



Conclusions ***:***

La valeur du coefficient de corrélation est égale à r = 0,9998. La valeur du coefficient de corrélation significative à 5%, avec un nombre de degré de liberté de 4, est de 0,811. Le r calculé est donc supérieur. Nous pouvons conclure que nous avons une corrélation linéaire qui est significative avec un risque d'erreur de 5%, c'est-à-dire que l'évolution des résultats calculés localement est similaire à celle des résultats fournis par l'éditeur.

La droite de régression linéaire a pour équation Y = 0,9981X + 1,5088 avec Y correspondant aux résultats calculés localement et X aux résultats fournis par l'éditeur. La pente de cette droite est de 0,9981. Nous pouvons conclure que les données locales varient peu par rapport aux données de l'éditeur.

Le compteur « recherche » de ACS donne localement de bons résultats.

Interprétations:

Les items utilisés localement pour repérer les recherches sont vraisemblablement les mêmes que ceux employés par l'éditeur pour ce compteur. Nous pouvons également supposer que les données sources utilisées sont similaires.

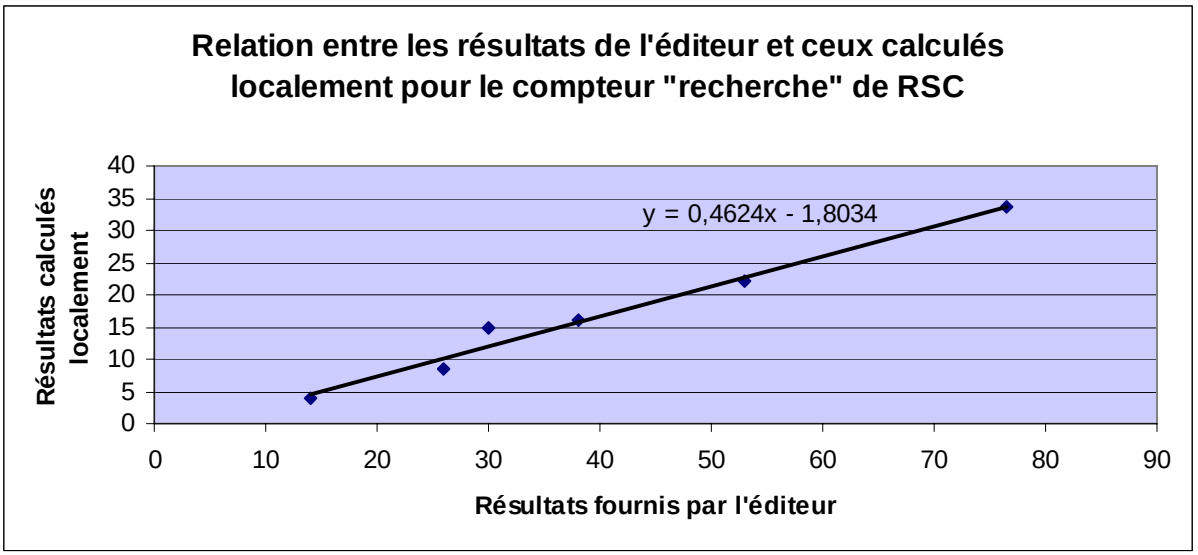
- *Deuxième situation*

Exemple : Compteur « recherche » de l'éditeur Royal Society of Chemistry

Résultats :

RSC	Recherche	
Mois	Résultats fournis par l'éditeur	Résultats calculés localement
Août 2006	26	9
Septembre 2006	30	15
Octobre 2006	77	34
Novembre 2006	38	16
Décembre 2006	14	4
Janvier 2007	53	22

Valeur du **coefficient de corrélation** : r = 0,9884



Conclusions :

La valeur du coefficient de corrélation est égale à r = 0,9884. La valeur du coefficient de corrélation significative à 5%, avec un nombre de degré de liberté de 4, est de 0,811. Le r calculé est donc supérieur. Nous pouvons donc conclure que nous avons une corrélation linéaire qui est significative avec un risque d'erreur de 5%, c'est-à-dire que l'évolution des résultats calculés localement est similaire à celle des résultats fournis par l'éditeur.

La droite de régression linéaire a pour équation Y = 0,4624X – 1,8034 avec Y correspondant aux résultats calculés localement et X aux résultats fournis par l'éditeur. La pente de cette droite est de 0,4624. Nous pouvons conclure que les données de l'éditeur sont supérieures à celles calculées localement. Elles sont majorées dans un rapport du simple au double.

Le compteur « recherche » de RSC évolue de façon satisfaisante mais ne donne pas localement de bons résultats.

Interprétations:

Les mesures locales sont inférieures à celles données par les éditeurs mais évoluent de façon similaire. On peut donc supposer qu'une seule partie des items mesurés par l'éditeur est mesurée au niveau local et que les items manquants et à identifier sont liés aux items déjà mesurés. Par exemple, dans le cas d'un compteur « recherche », l'éditeur pourrait mesurer à la fois les affichages des formulaires de recherche et les résultats des requêtes, ce qui est une erreur si on se réfère aux recommandations COUNTER.

N.B. Cette solution a été envisagée pour ce compteur mais les tests ne l'ont pas confirmée.

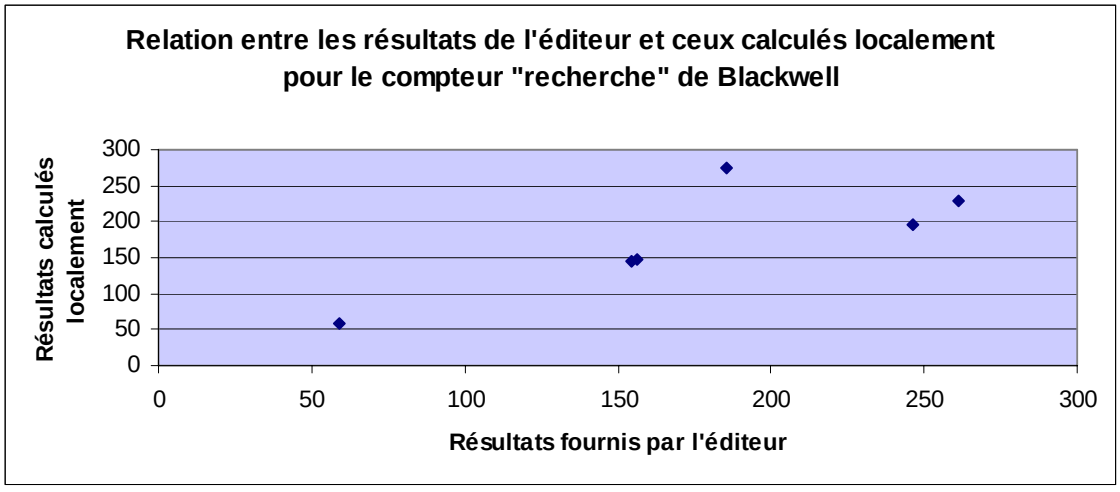
- *Troisième situation*

Exemple 1 : Compteur « recherche » de l'éditeur Blackwell

Résultats :

Blackwell	Recherche	
Mois	Résultats fournis par l'éditeur	Résultats calculés localement
Août 2006	59	60
Septembre 2006	155	145
Octobre 2006	261	229
Novembre 2006	247	195
Décembre 2006	157	147
Janvier 2007	186	274

Valeur du **coefficient de corrélation** : r = 0,7875



Conclusions :

La valeur du coefficient de corrélation est égale à r = 0,7875. La valeur du coefficient de corrélation significative à 5%, avec un nombre de degré de liberté de 4, est de 0,811. Le r calculé est donc inférieur. Nous pouvons donc conclure que nous avons une corrélation linéaire qui n'est pas significative avec un risque d'erreur de 5%, c'est-à-dire que l'évolution des résultats calculés localement est différente de celle des résultats fournis par l'éditeur.

Toutefois, l'observation du graphique représentant la relation entre les résultats de l'éditeur et ceux calculés localement montre qu'une seule donnée est responsable de cette corrélation modérée. Ce point correspond aux données du mois de janvier. Si l'on calcule le coefficient de corrélation des données du mois d'août au mois de décembre, nous avons pour résultat r = 0,9877 ce qui correspond à une forte corrélation. L'équation de la droite de régression linéaire tracée sur les données de la période d'août à décembre est la suivante, Y=0,7703X + 19,617. La pente de cette droite est de 0,7703.

Nous pouvons conclure que les données de l'éditeur sont supérieures à celles calculées localement. On peut tout de même les considérer comme proches les unes des autres

Interprétations:

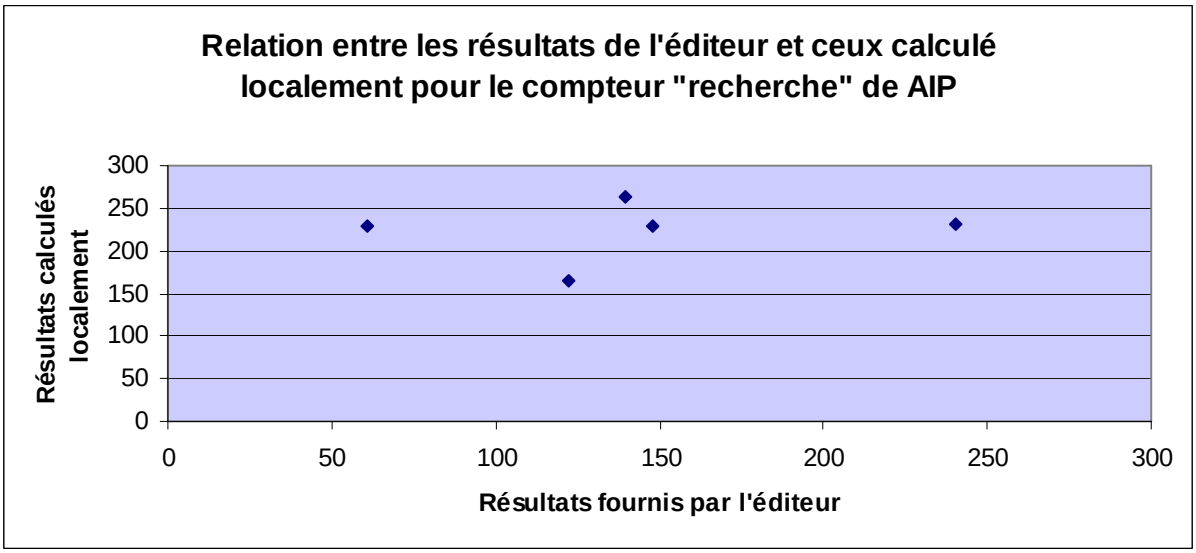
Au niveau local, le compteur a été calculé avec la même méthode du mois d'août au mois de janvier. Deux explications sont donc possibles au niveau de l'éditeur. Soit celui-ci a fourni une valeur erronée au mois de janvier, soit son système de mesure a changé.

Exemple 2 : Compteur « recherche » de l'éditeur American Institute of Physic

Résultats :

AIP	Recherche	
Mois	Résultats fournis par l'éditeur	Résultats calculés localement
Août 2006	123	164
Septembre 2006	241	231
Octobre 2006	140	264
Novembre 2006	148	230
Décembre 2006	61	230

Valeur du **coefficient de corrélation** : r = 0,1387



Conclusions :

La valeur du coefficient de corrélation est égale à r = 0,1387. La valeur du coefficient de corrélation significative à 5%, avec un nombre de degré de liberté de 3, est de 0,878. Le r calculé est donc inférieur. Nous pouvons donc conclure que nous avons une corrélation linéaire qui est faible, c'est-à-dire que l'évolution des résultats calculés localement est différente de celle des résultats fournis par l'éditeur.

Interprétations:

L'explication peut se situer aussi bien localement qu'au niveau de l'éditeur. Les items pris en compte sont vraisemblablement différents. Le problème peut également venir des données sources. Pourtant les rapports fournis par AIP suivent la recommandation COUNTER et au niveau local, les données sources ont été préparées et les compteurs mis en place en respectant également la recommandation, à l'exception de quelques limites. Toutefois, il faut remarquer que la plate-forme de l'éditeur AIP n'a pas été suffisamment étudiée pour s'assurer de la fiabilité du compteur.

Comparatif données Comité de suivi/rapports Counter établissements et consortium

	bases de données bibliographiques	autres bases de données	périodiques électroniques et bases de données de périodiques en texte intégral payants	livres électroniques et autres documents numériques dt ouv de référence	bases de données bibliographiques	autres bases de données	périodiques électroniques et bases de données de périodiques en texte intégral payants	livres électroniques et autres documents numériques dt ouv de référence
nombre de sessions par an	x	x	x		C	C		
nombre de sessions par mois	x	x	x		C	C		
nombre de sessions par an par catégorie d'usagers à desservir	x	x	x					
nombre de sessions par mois par catégorie d'usagers à desservir	x	x	x					
nombre de sessions par an et par titre	x	x	x		C	C		C
nombre de sessions par an et par service	x	x	x		C	C		C
nombre de sessions par mois et par titre	x	x	x		C	C		C
nombre de sessions par mois et par service	x	x	x		C	C		
nombre de sessions rejetées par an	x	x	x		C	C	C	C
nombre de sessions rejetées par an et par titre	x	x	x		C	C	C	C
nombre de sessions rejetées par mois	x	x	x		C	C	C	C
nombre de sessions rejetées par mois et par titre	x	x	x		C	C	C	C
nombre de recherches (requête intellectuelle) par an	x	x	x	x	C	C		C
nombre de recherches (requête intellectuelle) par an et par titre	x	x	x	x	C	C		C
nombre de recherches (requête intellectuelle) par an et par service	x	x	x	x	C	C	C	C
nombre de recherches (requête intellectuelle) par mois	x	x	x	x	C	C		C
nombre de recherches (requête intellectuelle) par mois et par titre	x	x	x	x	C	C		C
nombre de recherches (requête intellectuelle) par mois et par service	x	x	x	x	C	C	C	C
nombre de recherches (requête intellectuelle) par an et par catégorie d'usagers à desservir	x	x	x	x				
nombre de recherches (requête intellectuelle) par mois et par catégorie d'usagers à desservir	x	x	x	x				
nombre d'unités de contenu documentaire téléchargées par an		x	x	x			C	C
nombre d'unités de contenu documentaire téléchargées par mois		x	x	x			C	C
nombre d'unités de contenu documentaire téléchargées par an et par titre		x	x	x			C	C
nombre d'unités de contenu documentaire téléchargées par mois et par titre		x	x	x			C	C
nombre d'unités de contenu documentaire téléchargées par an et par catégorie d'usagers à desservir		x	x	x				
nombre d'unités de contenu documentaire téléchargées par mois et par catégorie d'usagers à desservir		x	x	x				
nombre d'unités de contenu documentaire téléchargées par type de page par an		x	x				C	
nombre d'unités de contenu documentaire téléchargées par type de page par an et par titre		x	x				C	
nombre d'unités de contenu documentaire téléchargées par type de page par mois		x	x				C	
nombre d'unités de contenu documentaire téléchargées par type de page par mois et par titre		x	x				C	
nombre d'unités de contenu documentaire téléchargées par type de page par an et par catégorie d'usagers à desservir		x	x					

Comparatif données Comité de suivi/rapports Counter établissements et consortium

nombre d'unités de contenu documentaire téléchargées par type de page par mois et par catégorie d'usagers à desservir			x	x					
nombre d'enregistrements téléchargés par an		x	x						
nombre d'enregistrements téléchargés par mois		x	x						
nombre d'enregistrements téléchargés par an et par titre		x	x						
nombre d'enregistrements téléchargés par mois et par titre		x	x						
nombre d'enregistrements téléchargés par an et par catégorie d'usagers à desservir		x	x						
nombre d'enregistrements téléchargés par mois et par catégorie d'usagers à desservir		x	x						

unité de contenu documentaire=article pour revues et chapitre ou section pour livres

C=Counter non obligatoire

C=Counter

C=rapports obligatoires pour consortia (les fournisseurs sont invités à fournir également JR2 DR2 DR3)

service=groupe de produits d'information en ligne protégé par une marque provenant d'un ou plusieurs fournisseurs, pour lequel on peut prendre un abonnement ou une licence et dont tout ou partie de la collection peut être interrogé

rapports locaux(tests) créés totalement ou partiellement, n'apparaissent pas les rapports supplémentaires non présents sur l'annexe 9 (voir annexe 13)

Ministère

Etablissement

SCD

Couperin

	Statistiques locales	Statistiques COUNTER
Avantages	Possibilité d'obtenir plus de résultats (62% des éditeurs, plates-formes… fournissant des statistiques non-Counter sur 108 ayant des statistiques) Possibilité d'obtenir des résultats précis (nombre d'articles courants, nombre d'archives, nombre d'articles par date de publication…) Homogénéité des tableaux de résultats et synthèse immédiate Souplesse dans le choix des rapports (mise en forme des données) Conformité aux besoins et objectifs du ministère et des établissements Possibilité de contrôler les statistiques des fournisseurs Statistiques par catégories d'utilisateurs plus aisées à obtenir Possibilité de construire son propre modèle économique	Homogénéité des données statistiques puisque respectant un standard Projet d'un officier de liaison ISO chez COUNTER Contrôle des statistiques des fournisseurs via les audits (rien avant juin 2007)
Inconvénients	Tests insuffisants à ce jour pour conclure ou non à une faisabilité avérée dans tous les domaines de la connaissance et pour tout type de document ; nécessité d'une vérification au cas par cas Interprétation des données difficile car hétérogénéité possible entre les éditeurs, pas de définition commune Difficulté pour mesurer l'usage de ressources gratuites Maintenance importante de l'outil (évolution des scripts en fonction des nouveaux fournisseurs et/ou des modifications sur leurs logiciels) Mise en place d'un dispositif technique indispensable (Proxy paramétré selon directives, PCs configurés pour passer par le proxy pour la documentation électronique…) Mutualisation nécessaire pour avoir résultats pour les consortia Pas de compatibilité avec Shibboleth aujourd'hui	26% seulement des éditeurs, plates-formes… sur 144 concernés par l'enquête sont COUNTER Aucun éditeur français labellisé à ce jour Documentation des fournisseurs insuffisante Malgré les directives, hétérogénéité possible si logiciels des plates-formes différents Mise en forme lourde ; nécessité d'inclure un outil de traitement automatique (extraction, synthèse) Erreuers dans les statistiques ; reprises nécessaires des résultats Standard national. Seulement 2 organismes français dans les membres et donc peu de possibilités de faire évoluer le standard Changements de plates-formes entraînant des "bugs" Nécessité de stocker toutes les données localement (fusion d'éditeurs…) Insuffisant pour les consortia