Les statistiques locales des ressources en ligne

Géraldine Barron, Claire Chédot-Leduc, Hélène Prost, Joachim Schöpfel

Les statistiques locales d'utilisation des ressources en ligne

Géraldine Barron, Claire Chédot-Leduc, Hélène Prost, Joachim Schöpfel

Dans I2D - Information, données & documents 2016/4 (Volume 53), pages 16 à 18
Éditions A.D.B.S.

ISSN 2428-2111
DOI 10.3917/i2d.164.0016

Les statistiques locales d'utilisation des ressources en ligne

[dispositif] Pour mesurer l'usage des ressources en ligne, à côté des statistiques des éditeurs, l'exploitation des statistiques locales produites par un serveur proxy est tout aussi intéressante.

L'analyse de l'usage des ressources en ligne est indispensable[1] pour comprendre le comportement des usagers et mener une politique documentaire raisonnée[2]. À côté des statistiques fournies par les éditeurs ou agrégateurs, on peut exploiter des statistiques locales, moins connues. Il y a cinq ans, seule une BU sur dix avait recours aux statistiques locales. Les avis sont partagés. Pour les uns, cette analyse représente une charge supplémentaire, une activité chronophage qui demande un soutien informatique. Pour d'autres, il s'agit d'une réelle alternative aux statistiques éditeurs avec l'avantage d'être contrôlées par l'institution. Voici quelques éléments pour mieux comprendre.

La fonction d'un proxy

Un établissement ou organisme peut générer ses propres statistiques à condition d'avoir installé un outil intermédiaire qui agit comme un filtre entre le lecteur et les ressources en ligne. Ce filtre, c'est le serveur proxy, aussi appelé proxy http ou proxy web. Au lieu d'accéder directement à la ressource, l'ordinateur du lecteur se connecte d'abord au serveur proxy qui va chercher les pages demandées avant de les renvoyer à l'utilisateur final. Cette procédure présente plusieurs avantages : comme seule l'adresse IP du proxy est « vue » par les sites Internet, la navigation est quasiment anonyme (y compris pour la géolocalisation), l'ordinateur du lecteur est mieux protégé et le serveur proxy autorise les accès en fonction des licences. C'est surtout ce dernier point qui rend le serveur proxy intéressant et incontournable pour les bibliothèques : les utilisateurs sont reconnus par leur adresse IP lorsqu'ils se connectent depuis un PC connu du serveur ou s'identifient par le biais d'un service d'authentification ; le serveur proxy leur ouvre alors l'accès à la ressource, y compris lorsqu'ils se connectent hors du campus. La gestion des accès par le biais d'un annuaire permet d'associer données de consultation et données personnelles de l'usager : type de lecteur (étudiant, enseignant, etc.), unité de rattachement (UFR, filière, service, laboratoire), niveau (LMD), etc., tout en s'assurant que les procédures réglementaires sont respectées (Cnil).

Les connexions au serveur proxy laissent des « traces » et génèrent des fichiers log[3] avec enregistrement de l'historique des événements : qui s'est connecté, quand, pour accéder à quel site et page (sommaire, résumé, plein texte, etc.), avec quel résultat (visualisation, téléchargement), etc. Notre exemple (figure 1) est une ligne extraite d'un fichier log qui identifie l'IP utilisateur (194.57.180.3), le login utilisateur (dup003), la date de la requête (12 juin 2013), l'éditeur (Lextenso), la page demandée[4] et le code retour (200) qui indique que l'article demandé a été envoyé à l'usager. L'exploitation de ces enregistrements produit des statistiques locales. Le problème est que les fichiers log sont différents pour chaque éditeur, qu'ils changent avec le temps et qu'il faut les décrypter à l'aide d'un programme (« parseur ») à mettre régulièrement à jour.

Comparaison des statistiques locales et éditeurs

L'analyse des fichiers log présente le même intérêt que les statistiques Counter[5] (voir encadré), à savoir que l'on définit clairement la valeur d'une « session » ou d'une « requête ». En effet, une statistique n'a de valeur en soi que dans une démarche comparative : des statistiques normalisées permettent de comparer entre elles les sessions sur différentes ressources et de voir évoluer les consultations d'une même ressource dans le temps. Puisque tous les éditeurs ne respectent pas la norme Counter, seule l'analyse des logs permet de mettre en regard les statistiques de l'ensemble des abonnements d'un établissement. À condition qu'un parseur ait été rédigé pour décrypter les fichiers log de chaque éditeur, ce que le consortium Couperin s'emploie à réaliser de manière mutualisée[6].

Figure 1

Exemple de ligne d'un fichier log

```
194.57.180.3 - dup003 [12/Jun/2013:15:43:15 +0100] «GET http://www.lextenso.fr/weblextenso/article/afficher?id=C010IXCXCX2001X12X01X00177X053&origin=recherche&d=3575204329777 HTTP/1.1» 200
```

1. C. Chédot-Leduc, G. Barron. « Université. Une collaboration pleine de ressources ». *I2D – Information, données et documents*, 2015, n° 2, p.7-9

2. C. Boukacem-Zeghmouri (coord.). *L'information scientifique et technique dans l'univers numérique. Mesures et usages*. ADBS éditions, 2010. C. Boukacem-Zeghmouri. « Résultats de l'enquête Couperin sur l'utilisation des statistiques de consultation en BU ». *In : Journée d'étude Couperin sur les statistiques d'utilisation des ressources électroniques*, Paris, 23 mars 2012, www.couperin.org/images/stories/documents/Statistiques/JE_23_MARS_2012/prsent_enqute_couperin_23_mars_cbz-1.pdf

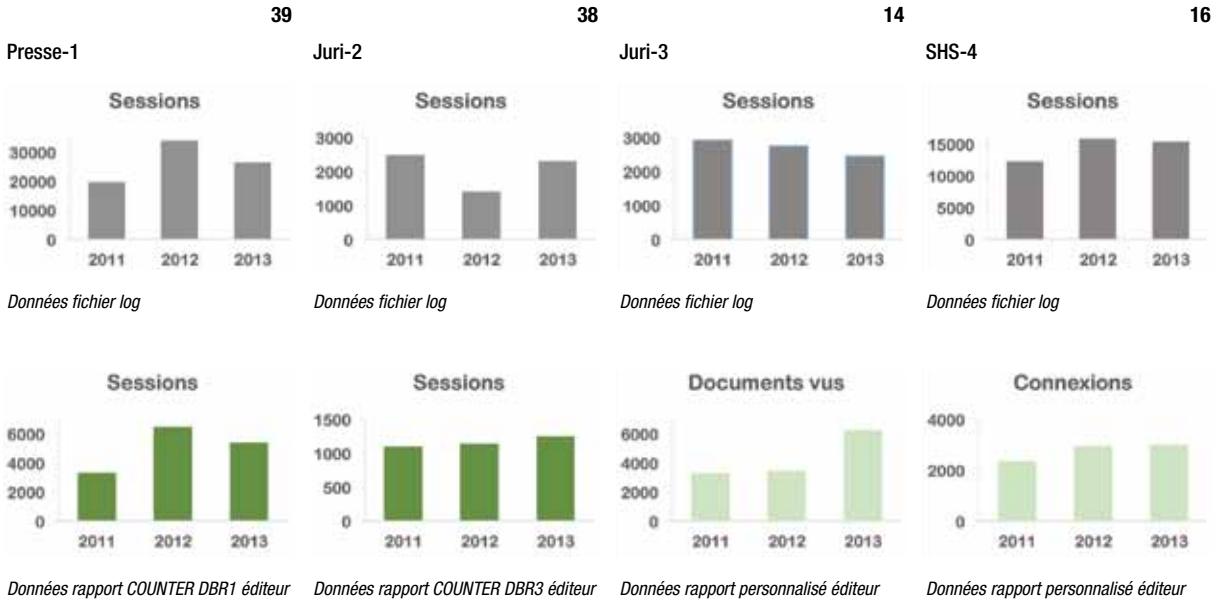
3. Fichiers journal ou de journalisation

4. www.lextenso.fr/weblextenso/article/afficher?id=C010IXCXCX2001X12X01X00177X053

5. C. Boukacem-Zeghmouri, J. Schöpfel. « Statistiques d'utilisation des ressources électroniques en ligne : le projet Counter ». *Bulletin des Bibliothèques de France*, 2005, n° 4, p. 62-66

6. http://ang.couperin.org

Figure 2
Statistiques locales et éditeurs, extraites d'une étude sur les usages de ressources électroniques de la Bulco[1]



1. C. Chédot-Leduc, G. Barron. *Usage des ressources électroniques de la Bibliothèque universitaire du Littoral Côte d'Opale (URElecBULCO) : rapport d'étude.* Université du Littoral Côte d'Opale, Dunkerque, 2014. https://hal.archives-ouvertes.fr/hal-01367939

Bien que les statistiques fournies par les éditeurs et celles produites à partir des fichiers log donnent des résultats différents pour des raisons techniques (comptage de clics) ou d'interprétation, l'étude des corrélations entre les deux séries de valeurs montre que les profils de consultation sont généralement assez similaires[7]. Toutefois, la réalité peut être différente, comme le montre la figure 2.

La première ligne montre les statistiques locales (sessions), la deuxième celles délivrées par les éditeurs ou agrégateurs (sessions, documents vus, connexions). Chaque colonne correspond à un éditeur ou agrégateur (de gauche à droite : Presse-1, base de presse en texte intégral conforme Counter ; Juri-2, base de données juridique conforme Counter ; Juri-3, base de données juridique ; SHS-4, bouquet de revues SHS). Plusieurs constats s'imposent :

- que les rapports des éditeurs soient Counter ou non, qu'ils fournissent une mesure similaire à celle des fichiers log (session/connexion) ou non (documents vus), il est vain de vouloir les comparer avec les statistiques locales ;
- les chiffres sont assez différents (*cf.* Juri-3), les statistiques locales sont souvent supérieures ; mais attention, ces chiffres ne comptent pas nécessairement les mêmes événements ;
- parfois, les tendances sont similaires (Presse-1, SHS-4), parfois différentes (Juri-2) voire contraires (Juri-3) ;
- seule la comparaison horizontale des statistiques locales est possible puisque la mesure est la même quel que soit l'éditeur ; / / / /

7. J. Duy, L. Vaughan. « Usage data for electronic resources: A comparison between locally collected and vendor-provided statistics ». *The Journal of Academic Librarianship*, 2003, n° 29, p. 16-22

« ezPaarse et ezMesure »

> **ezPaarse**[1] est un progiciel d'analyse des accès aux ressources électroniques qui transforme les fichiers log en statistiques ou « évènements de consultation ». Il permet l'analyse, l'enrichissement et l'exploitation des logs d'accès. ezPaarse se présente sous la forme d'une application web disposant d'un formulaire et d'une API permettant l'ingestion manuelle et la traduction automatique des logs générés par les serveurs et proxy des établissements, puis de les délivrer sous forme de fichier propre au format csv. Cette exploitation s'appuie sur des programmes appelés parseurs qui découpent et traduisent les lignes de logs. Regroupés dans la plateforme AnalogIST[2], ces parseurs sont maintenus et régulièrement mis à jour par la communauté des utilisateurs d'ezPaarse, si des modifications sont repérées sur le site de l'éditeur.

> Le projet **ezMesure**[3] souhaite « *agréger, comparer, visualiser, valoriser* » les statistiques locales; il a l'ambition d'être l'agrégateur national des statistiques locales produites par les instances d'ezPaarse installées dans les établissements. ezMesure propose une interface en ligne de visualisation dynamique, consolidée et comparative des données. Cet agrégateur complétera le portail Mesure existant, qui assure déjà l'agrégation des statistiques fournies par les éditeurs. ■

1. http://ezpaarse.couperin.org
2. http://analogist.couperin.org
3. www.couperin.org/groupes-de-travail-et-projets-deap/statistiques-dusage/ezmesure

//// on peut donc constater que Presse-1 et SHS-4 ont attiré davantage de sessions que Juri-2 et Juri-3 et que la consultation de Juri-3 baisse tandis qu'elle augmente pour SHS-4 ;

- les statistiques fournies par les éditeurs renseignent essentiellement sur l'évolution de la ressource d'une période à l'autre.

Si le proxy est géré par la bibliothèque, l'obtention des fichiers log est relativement simple. Par contre, traduire les fichiers log en événements de consultation (statistiques) peut être un travail fastidieux qui demande des compétences techniques et des outils pointus. *A priori*, il y a 4 options : utiliser des scripts Perl ou Python pour extraire les données ; faire appel à des logiciels d'analyse de fichiers log comme Sawmill, AWStats ou AnalogX ; convertir les fichiers log en CSV avec un logiciel analyseur (Stream Editor, Awk) ; ou avoir recours à l'outil *open source* ezPaarse (voir encadré) pour transformer les fichiers log en statistiques de consultation. Le projet Mesure (voir encadré) de Couperin et l'Inist (CNRS) met à disposition la plateforme AnalogIST qui permet d'externaliser ce travail[8].

Fiabilité et limites des statistiques locales

Pourquoi ces différences avec les statistiques éditeurs ? Les transactions ne passent pas toutes par le proxy et donc certaines transactions ne laissent pas de trace dans les fichiers log. Pour des raisons de confort d'utilisation, de nombreux établissements déclarent des plages d'adresses IP autorisées qui couvrent leur serveur proxy mais également leur parc informatique. Si un utilisateur se connecte depuis un PC de l'établissement *via* le service documentaire, le lien vers la ressource sera « proxyfié » : on ne saura pas qui s'est connecté mais on connaîtra l'IP, donc le PC et le bâtiment, service, etc. où il se trouve. En revanche, si l'utilisateur est dirigé vers un article depuis un moteur de recherche, l'accès à la ressource sera « transparent » pour l'utilisateur comme pour l'établissement : le cheminement vers la ressource ne passera pas par le proxy et l'utilisateur, reconnu comme autorisé par son IP, n'aura pas à s'identifier[9]. Notons enfin que les annuaires, comportant souvent une partie déclarative, ne sont pas toujours parfaitement renseignés, ce qui constitue une limite dans l'identification de profils d'utilisateurs.

Malgré ces limites, l'intérêt majeur des statistiques locales est de présenter de manière unifiée et fine les statistiques d'usage des ressources numériques. Elles pallient la variété ou l'absence des statistiques éditeurs et donnent une vision plus homogène de l'usage de ces ressources. En principe, ces statistiques peuvent aussi couvrir les ressources en accès libre (archives ouvertes, bases de revues en libre accès, entrepôts de données), à condition que l'usager utilise les outils de la bibliothèque pour s'y connecter.

Soulignons qu'il est difficile d'interpréter la valeur d'un clic, d'une requête ou d'une session ; les indicateurs tels que le coût à la connexion, le coût au téléchargement ou le nombre moyen d'accès/de téléchargements par utilisateur révèlent une tendance, ils ne sont pas le reflet fidèle des intentions et comportements. En guise de conclusion, trois remarques : les statistiques locales et éditeurs sont complémentaires mais pas interchangeables ; elles ne sont pas fiables à 100 % ; et, pour mieux connaître l'usage en ligne, il faut (aussi) passer par des enquêtes qualitatives. ■

> **Géraldine Barron**
Conservatrice, Bibliothèque de l'Université du Littoral Côte d'Opale
Geraldine.Barron@univ-littoral.fr

> **Claire Chédot-Leduc**
Maître de conférences, Université Littoral Côte d'Opale
Claire.Leduc@univ-littoral.fr

> **Hélène Prost**
Membre associé GERiiCO
helene.prost007@gmail.com

> **Joachim Schöpfel**
Maître de conférences, Université de Lille
joachim.schopfel@univ-lille3.fr

8. M. Joly, T. Porquet, G. Colcanap. « Quantifier les accès à la documentation numérique payante : un panorama des enjeux » et D. Lechaudel, S. Gully. « ezPAARSE : outil d'évaluation de l'usage des ressources numériques ». *In* : *L'état de l'art des réseaux de l'enseignement et de la recherche*, JRES, Montpellier, 12 décembre 2013, https://2013.jres.org

9. On peut partiellement contourner ce problème si la bibliothèque se fait référencer par Google Scholar.

À propos des statistiques éditeurs

Les éditeurs fournissent des données relatives aux différentes requêtes enregistrées sur leur plateforme et comptabilisent le nombre d'envois d'articles effectués depuis leur serveur. Ces statistiques sont plus le reflet de l'activité sur la plateforme que de l'usage d'un bouquet défini de revues. Elles ne contiennent aucune information sur l'utilisateur. Si ces statistiques sont accessibles sur le site de l'éditeur, leur collecte peut être ardue et chronophage pour les bibliothécaires. Se présentant sous des formats variés, il est souvent nécessaire de retravailler les fichiers pour les exploiter ensemble, même s'ils sont conformes au code de bonnes pratiques Counter[1]. Créée sur l'initiative d'un consortium d'experts des ressources électroniques, ce code a l'avantage de définir dans un langage précis et facilement compréhensible les statistiques à fournir pour illustrer de façon exhaustive la consultation de l'ensemble des ressources électroniques (revues, *ebooks*, bases de données, médias). La version 4 du code Counter propose 24 rapports de statistiques, dont 8 se rapportent aux revues.

Pour automatiser la collecte des statistiques conformes à Counter, le consortium Couperin a mis en place la plateforme Mutualisation et évaluation des statistiques d'utilisation des ressources électroniques (Mesure)[2]. Ce portail moissonne régulièrement les rapports statistiques JR1 et JR1a, à l'aide du protocole Sushi (Standardized Usage Statistics Harvesting Initiative) et présente les données sous forme de tableaux et graphiques. Ces données ainsi récupérées restent consultables et fournissent un historique des usages. ■

1. https://www.projectcounter.org
2. www.couperin.org/groupes-de-travail-et-projets-deap/statistiques-dusage/mesure