

Quantifier les accès à la documentation numérique payante : un panorama des enjeux

Monique Joly

Directrice SCD Doc'INSA - INSA de Lyon
Bibliothèque Marie Curie / Campus Lyon Tech
31 avenue Jean Capelle
69621 Villeurbanne

Grégory Colcanap

Directeur du SCD de l'Université d'Evry - Val d'Essonne
2 rue André Lalande
91025 EVRY CEDEX

Thomas Porquet

Département Etudes et Prospective - Consortium Couperin
SCD de l'Université Paris Descartes
45 rue des Saints Pères
75006 PARIS

Résumé

Un panorama synthétique des enjeux politiques, stratégiques et financiers de l'accès à la documentation électronique payante sera dressé pour appuyer la présentation du dispositif ezPAARSE/AnalogIST lancé conjointement par l'INIST-CNRS, le consortium Couperin et l'Université de Lorraine.

Le volume financier de la documentation numérique est en constante augmentation alors que les établissements membres de Couperin déclarent depuis peu une baisse de leurs budgets consacrés à la documentation. Nous avons donc sans cesse besoin d'arguments pour continuer de négocier la modération des tarifs.

Dans ce contexte, nous soulignerons la nécessité stratégique de disposer de données de consultation fiables, catégorisées et produites à partir de données locales pour que les établissements de l'ESR prennent, ou reprennent, le contrôle du suivi de leurs usages documentaires.

Cette réappropriation passe bien sûr par la prise de conscience des enjeux et dépend de l'adoption de solutions techniques compatibles. Nous proposons l'adoption large du progiciel ezPAARSE et la mutualisation des connaissances dans le réseau AnalogIST.

Mots-clefs

documentation numérique, accès, mesurer, panorama, mutualisation, usages, politique documentaire, dsi, scd, pilotage, politique documentaire, fichier log, reverse-proxy

1 Note préliminaire

Grégory Colcanap et Monique Joly sont respectivement le coordinateur et la responsable du département Services et Prospective du consortium Couperin.org¹. Thomas Porquet est chargé de mission pour le département « Services et Prospective ».

Le consortium existe depuis 1999 pour évaluer, négocier et organiser l'achat ou l'abonnement de ressources documentaires numériques. Il organise également la veille autour de la diffusion des publications numériques et

1 <http://www.couperin.org/>

développe des services pour ses membres notamment dans le traitement de leurs statistiques. Il réunit aujourd'hui 228 établissements membres et anime un réseau d'environ 70 négociateurs pour plus de 200 négociations.

2 Introduction

Nous souhaitons présenter d'abord un court panorama des enjeux politiques, stratégiques et financiers de l'accès à la documentation numérique payante. Nous détaillons ensuite les arguments qui parlent en faveur d'une réappropriation par les établissements de leurs statistiques de consultation. Puis nous présentons la solution de l'analyse de fichiers de traces localement collectés pour la constitution de statistiques d'accès.

Tous ces éléments peignent le contexte dans lequel le dispositif ezPAARSE/AnalogIST s'inscrit, ce qui en est attendu et comment il s'organise.

3 Le panorama général des enjeux politiques, stratégiques et financiers de l'accès à la documentation numérique payante

Le contexte très général est celui de l'Information et de la documentation Scientifique et Technique¹ qui « s'inscrivent, à l'instar d'autres outils, au cœur du processus de recherche et d'enseignement. »

C'est plus précisément du volet documentation numérique payante, partie constitutive de l'IST, dont il sera question ici.

Commençons par fournir quelques ordres de grandeurs :

- Le chiffre d'affaires de l'IST mondial représente environ 25 milliards de dollars et connaît une croissance de 4 à 5 % par an depuis 2007.
- Ce chiffre d'affaires est réalisé majoritairement aux États-Unis (environ 52 %) et pour un tiers en Europe et au Moyen-Orient.
- Le nombre d'éditeurs est estimé entre 5000 et 10000. 28000 revues à comité de lecture sont diffusées en ligne et publient plus de 1,8 millions d'articles/an.
- 1,5 milliard d'articles sont téléchargés, par environ 10 millions d'utilisateurs, chaque année
- En France, en 2012, plus de 70 millions d'euros sont dépensés par les établissements de l'enseignement supérieur et de la recherche pour la documentation électronique²
- Dans le même temps, ces mêmes établissements déclarent disposer de budgets en forte baisse pour l'achat de la documentation (moins 9 % en moyenne).

Ainsi, cette documentation numérique est devenue centrale pour la Recherche et y accéder apporte un avantage incontestable aux communautés scientifiques, mais il y a une forte tension pour accéder à la ressource en raison de son coût. L'enjeu est d'autant plus important que l'observation des variations annuelles montre, sur 20 ans, une envolée des prix qui n'est plus supportable et qui conduit les établissements à stopper certains abonnements. Ce phénomène, que la profession nomme « serial crisis »³, conduit à la révision des politiques documentaires et à l'obligation de disposer d'indicateurs permettant par exemple de choisir les abonnements à interrompre.

L'utilisation⁴ fait bien entendu partie des critères retenus pour l'élaboration des indicateurs (coût par article accédé, coût par utilisateur réel, ...).

À ce jour les statistiques d'utilisation sont contractuellement obtenues, auprès des éditeurs, dans le cadre des négociations de Couperin.org. Pour plus de fiabilité et d'homogénéité, ces données doivent être validées et complétées par des données d'utilisation produites localement.

2 Enquête annuelle interne du Ministère de l'Enseignement Supérieur et de la Recherche et de Couperin.org.

3 Monograph & Serial Costs in ARL Libraries, 1986-2011 <http://www.arl.org/storage/documents/monograph-serial-costs.pdf>

4 Nous préférons ici parler d'utilisation plutôt que de consultation ou d'usage : le fait qu'un usager a eu accès à une revue ou un article nous permet de constater un intérêt mais ne nous permet pas vraiment de qualifier la consultation en elle-même.

Mieux piloter la politique documentaire à l'aide d'indicateurs maîtrisés constitue donc l'enjeu essentiel du projet ezPAARSE. Détaillons plus largement, ci-après les limites du système actuel et les évolutions souhaitées.

4 Pourquoi vouloir quantifier les accès à la documentation numérique

Les statistiques d'utilisation des ressources numériques nous intéressent à plusieurs titres :

- Elles contribuent à **orienter la stratégie des négociations tarifaires** pour le consortium. La connaissance fine des besoins et des usages dégagent des arguments utiles.
- Elles entrent dans la définition des **politiques d'acquisition** : priorités, choix, meilleure connaissance des besoins des publics par observation des utilisations réelles.
- Elles soutiennent des **politiques de services** et d'accompagnement des publics vers l'usage documentaire. C'est le cas par exemple quand nous constatons que certains publics n'utilisent pas ou peu une source pertinente pour eux.
- Elles permettent le ciblage des publics pour évaluer leur satisfaction d'un outil documentaire qu'ils utilisent.

4.1 Les limites des statistiques fournies par les éditeurs

4.1.1 Difficulté de se fier, sans réserve, aux données communiquées

Même s'ils sont de plus en plus nombreux à s'y prêter, tous les éditeurs ne fournissent pas encore de statistiques d'utilisation de leurs contenus. Quand c'est le cas nous ne pouvons pas faire complètement confiance aux données qui nous parviennent si nous ne connaissons pas précisément les méthodes de comptage employées. En effet, ces statistiques ne sont pas toujours agrégées d'une façon publiquement documentée, ni fournies régulièrement, ni forcément dans un format de référence.

L'initiative de normalisation qui rencontre le plus de succès est celle du projet COUNTER⁵ qui préconise des bonnes pratiques et renseigne précisément les méthodes de comptage.

Mais, de façon générale, si nous voulons obtenir un **éclairage contradictoire** il nous semble crucial de pouvoir croiser les données des éditeurs avec les nôtres.

4.1.2 Difficulté d'obtenir les données communiquées

Il reste néanmoins malaisé de récupérer et travailler avec les statistiques fournies par les éditeurs. Il faut souvent les collecter manuellement sur le site de l'éditeur après s'être authentifié car le protocole de moissonnage par Webservice SUSHI⁶, qui permet l'automatisation de ce travail, n'est pas non plus systématiquement mis en œuvre. Le projet MESURE⁷ pour collecter et agréger automatiquement des données en provenance des éditeurs est un premier pas dans cette direction.

4.1.3 Difficulté de partager et comparer les données obtenues

Nous voulons pouvoir partager librement nos données, par exemple au sein du Couperin.org, entre les établissements, afin de disposer de vues globales cohérentes et homogènes. Les deux catégories de problèmes sus-mentionnés : méthodes de comptage et moissonnage, sont des freins non négligeables à la réalisation de cet objectif.

5 <http://www.projectcounter.org/>

6 http://www.niso.org/apps/group_public/project/details.php?project_id=111

7 <http://mesure.couperin.org/>

4.1.4 Manque d'information sur les types d'utilisateurs

Malgré les avancées indéniables qu'il a permises, le guide de bonnes pratiques COUNTER ne suffit pas à satisfaire notre besoin. En effet, les données fournies dans les rapports y sont agrégées à l'échelle de l'établissement et le détail des catégories d'utilisateurs n'y figure pas.

Par exemple : pour chaque titre de périodique présent sur la plateforme d'un éditeur, on saura que les utilisateurs d'un établissement ont eu accès aux articles un certain nombre de fois mais on ne saura rien du profil de ces utilisateurs. Sont-ils étudiants, de quel niveau ? Sont-ils enseignants, de quelle discipline ? Sont-ils chercheurs, etc...

4.2 L'analyse est rendue possible avec du matériel collecté localement

L'utilisation d'un logiciel de fédération d'identité⁸ pour accéder à distance aux ressources numériques sur le site d'un éditeur est attrayante mais des données complémentaires sur la personne présente dans les annuaires⁹ sont ainsi concédées, qui constituent une inestimable fuite d'informations stratégiques sur nos publics et permettent aux éditeurs de formaliser des offres plus coûteuses, encore, parce que segmentées. De plus, l'utilisation d'un tel système entraîne aussi la perte des traces des requêtes émises par l'utilisateur car l'activité ne se déroule plus qu'entre son navigateur et le site de l'éditeur une fois l'authentification réussie.

Pour ces deux raisons, nous préconisons actuellement, pour l'accès distant, la solution technique du reverse-proxy par lequel transitent tous les utilisateurs quand ils veulent accéder aux ressources souscrites par leur établissement, ce qui permet à la fois :

- De ne pas divulguer, aux éditeurs, d'attributs stratégiques sur nos utilisateurs.
- De collecter les traces de passage de l'utilisateur, inscrites à chacune de leurs actions dans un fichier de log.

L'obtention de ce matériel local est donc particulièrement stratégique en ce qu'il est porteur de très nombreuses informations que le système ezPAARSE, développé par un partenariat INIST-CNRS, Couperin.org et Université de Lorraine, pourra exploiter et élargir à d'autres partenaires.

4.3 Utiliser ezPAARSE et participer au réseau AnalogIST

Obtenir des données qui quantifient l'utilisation des ressources numériques à partir des fichiers de log d'un reverse-proxy nécessite :

- La mise en place d'une politique d'archivage pour les fichiers de log, que nous encourageons et soutenons en prenant contact avec les établissements ayant répondu à notre enquête en mars 2013 et déclaré l'utilisation d'un reverse-proxy pour gérer leurs accès distants, afin de leur indiquer la marche à suivre pour mener cette politique d'archivage
- La conduite de traitements réguliers (filtrage, enrichissement, qualification) que la brique logicielle ezPAARSE¹⁰ se charge d'effectuer. En développement depuis novembre 2012, ezPAARSE est désormais disponible dans une version stable et complète que l'on peut tester dans une version en ligne avant de l'installer localement.
- Une participation autant que possible au réseau AnalogIST¹¹, pensé comme une plateforme nationale et « un espace collaboratif permettant la mise en commun des analyses de plateformes d'éditeurs de littérature scientifique. » C'est cette participation qui garantit qu'ezPAARSE sera capable d'analyser et d'interpréter correctement, avec ses parseurs, les événements de consultation présents dans les fichiers de log.

8 Comme Shibboleth, voir [http://fr.wikipedia.org/wiki/Shibboleth_\(f%C3%A9d%C3%A9ration_d'identit%C3%A9\)](http://fr.wikipedia.org/wiki/Shibboleth_(f%C3%A9d%C3%A9ration_d'identit%C3%A9))

9 Sa catégorie, sa discipline, etc.

10 <http://ezparse.couperin.org>

11 <http://analogist.couperin.org>

Pour leur installation et leur bon fonctionnement dans les établissements intéressés, ces outils requièrent le concours des Directions des Systèmes d'Information en partenariat avec les Directions des Bibliothèques. L'équipe réunie autour d'ezPAARSE et d'AnalogIST propose son aide sur les différents aspects exposés.

Les résultats obtenus serviront à l'élaboration d'indicateurs utiles à la validation des politiques documentaires et revêtent en cela un caractère stratégique pour les établissements de l'ESR.

4.4 Conclusion

Le projet national de plateforme AnalogIST est lancé depuis novembre 2012 et son succès reposera sur la constitution d'un réseau pluridisciplinaire. Ce réseau assurera l'analyse et la veille des plate-formes éditeurs (pour l'actualisation des bases de connaissance éditeurs) et la maintenance des outils (parseurs).

La réalisation d'un travail commun sur ces questions, éminemment politiques, implique conjointement les DSI, les SCD et les directions des établissements. Nous espérons en avoir souligné l'importance, tant du point de vue de la stratégie interne à l'établissement (connaissance des usages pour piloter la politique d'acquisition et la politique des services documentaires) que du point de vue externe (négociations annuelles ou pluriannuelles d'abonnements à des ressources documentaires payantes) afin d'optimiser les investissements consentis par la puissance publique pour le développement de l'Enseignement Supérieur et de la Recherche.

Bibliographie

- [1] WARE Mark, MABE Michael. The STRM Report, An overview of scientific and scholarly journal publishing [en ligne]. International Association of Scientific, Technical and Medical Publishers 2012. http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf
- [2] COLCANAP Grégory, MALOTAUX Sandrine. Le coût de l'information scientifique et technique. Communiqué de presse, 3 septembre 2013. <http://www.couperin.org/images/stories/documents/Communique%20presse%20Cot%20de%20IIST%202013%2009%2003.pdf>
- [3] STEMPER James, JAGUSZEWSKI Janice. Usage Statistics for Electronic Journals. Collection Management, 28(4): 3–22, 2003. <http://purl.umn.edu/109732>
- [4] LOGRE Julien. Autour de COUNTER. L'utilisation des ressources électroniques onéreuses spécialisées dans l'enseignement supérieur et la recherche en France : mesures, indicateurs, pilotage [en ligne]. Enssib 2013. <http://www.enssib.fr/bibliotheque-numerique/notice-60371>
- [5] LETROUIT Carole, DAZY André. La problématique de l'accès distant [en ligne]. Couperin 2012. <http://www.couperin.org/groupes-de-travail-et-projets-deap/acces-aux-ressources-cat/acces-distant/131-acces-distant/644-lacces-distant-shibboleth-et-les-federations-didentites>