IFLA WLIC 2019 Athens

Submitted on: 16/09/2019

Title of the Satellite Meeting: Data intelligence in libraries: The actual and artificial perspectives
Date: August 22, 2019
Location: German National Library – Deutsche Nationalbibliothek (DNB), Frankfurt, Germany

ezPAARSE and ezMESURE: Assembling dashboards on a national repository from fine-grained and locally generated usage statistics to electronic resources

Thomas Porquet
Département Services et Prospective, Couperin.org, Paris, France.
thomas.porquet@couperin.org

Lechaudel Dominique
Département Projets et Innovation, Inist – CNRS, Vandoeuvre-lès-Nancy, France
dominique.lechaudel@inist.fr

Copyright © 2019 by Thomas Porquet and Dominique Lechaudel. This work is made available under the terms of the Creative Commons Attribution 4.0 International License:
http://creativecommons.org/licenses/by/4.0

Abstract:

Generating fine-grained Access Events (AEs) to electronic resources from locally gathered log files through ezPAARSE begins to be a fairly well known process among colleagues having to manage a reverse-proxy to provide their patrons with authenticated access to their institution's electronic resources' subscriptions. Our objectives with this software are manifold: let our users gain control over the generation of usage statistics; homogenize the production of this strategic data; have more complete and granular data than those usually provided by publishers, to supplement them (including information on categories of users) or to compare them.

ezPAARSE has been implemented as free and open source software since the end of 2012 and now comes with a list of more than 200 community contributed parsers, providing its international user base with a good coverage of their electronic resources subscriptions. For the publishers or providers that are not recognized by ezPAARSE yet or simply need an update, it is easy to contribute an analysis on our community platform. This is were the different types of URLs are collected, semi-automatically analyzed and commented so the corresponding parser can be accurate. We reward this participation with a system of open badges.

Collecting ezPAARSE generated and enriched AEs into the ezMESURE french national repository for dynamic dashboard creation, consolidation and representations has been an sustained effort since the

beginning of 2016 for INIST-CNRS, Couperin.org and its members. The repository now hosts access data for almost 60 institutions. Based on an ElasticSearch and Kibana stack, ezMESURE is able to store, retrieve and display all this data in a variety of dashboards that can be tweaked to every need while also paving the way for aggregated views. With this ecosystem, our users are provided with a complete processing chain, to finely analyze their data and get strategic elements for driving their subscription campaigns.

Keywords: usage data, electronic resources, visualization, dashboard, log analysis

---

1 INTRODUCTION

Dans cet article, nous passerons en revue les différents éléments qui constituent la chaîne de traitement mise à disposition des établissements membres du consortium Couperin.org en particulier pour la génération de leurs propres statistiques d'utilisation des ressources numériques auxquelles ils s'abonnent. Les objectifs poursuivis sont multiples : obtenir des données homogènes pour les éditeurs qui ne se conformeraient pas (encore) à COUNTER, et des données contradictoires pour les éditeurs qui s'y conforment [1], avoir accès à un niveau de détail que n'offrent pas les statistiques fournies par les éditeurs, proposer un dispositif qui permet le croisement pour plusieurs établissements, etc.

2 L'ANALYSE DE LOGS COMME PRINCIPE DE FONCTIONNEMENT SOUS-JACENT

2.1 Un contexte technique assez classique

L'analyse des fichiers journaux (appelés "logs" dans la suite de l'article)[1] est un processus courant, qu'il soit mis en place par un service informatique pour surveiller le bon fonctionnement d'applications logicielles ou par un service marketing qui souhaite comprendre comment les utilisateurs se comportent en ligne quand ils naviguent sur un site ou un ensemble de sites. De nombreux logiciels d'analyse web basée sur les logs existent[2] et on peut citer ici les plus connus : Awstats, Piwik (devenu Matomo), etc.

Même s'il existe d'autres dispositifs, ce principe (des logs et un outil qui les analyse) est souvent celui qui est appliqué dans le domaine des statistiques d'utilisation, lorsqu'il s'agit de mesurer la fréquence d'accès aux ressources électroniques auxquelles les établissements de l'enseignement supérieur et de la recherche s'abonnent. Tout comme les éditeurs et les fournisseurs de plateformes collectent des logs de leurs serveurs pour générer et fournir à leurs clients des rapports COUNTER[3], les institutions qui s'abonnent à des ressources et utilisent un serveur proxy pour authentifier leurs utilisateurs[4] peuvent elles-aussi analyser le trafic sortant pour en extraire des éléments à grain fin : les événements de consultation.

1 https://en.wikipedia.org/wiki/Log_analysis
2 https://en.wikipedia.org/wiki/List_of_web_analytics_software
3 https://www.projectcounter.org/code-of-practice-five-sections/6-logging-usage/
4 Le plus connu d'entre eux s'appelle EZproxy, commercialisé par OCLC, mais il en existe d'autres

2.2 ezPAARSE : une spécialisation aux ressources documentaires

2.2.1 Intérêt et mise en œuvre

Pour analyser précisément le trafic des usagers qui transitent par le proxy d'un établissement, le logiciel libre et gratuit ezPAARSE est proposé depuis 2013 [2]. Issu d'une collaboration étroite et prolifique entre le consortium de négociation documentaire Couperin.org [5] et l'INIST [6], unité propre de service du CNRS, et plusieurs établissements pilotes, dont l'Université de Lorraine.

Son intérêt principal réside dans le filtrage de tout ce qui n'est pas jugé pertinent et l'enrichissement des événements de consultation retenus avec des données bibliographiques explicites et précises à chaque fois que cela est possible.
D'installation relativement aisée et rapide, ezPAARSE se déploie et peut s'utiliser ponctuellement ou s'automatiser pour être déclenché à fréquence régulière pour traiter les logs du jour, de la semaine, du mois, etc. C'est une tâche ponctuelle qui incombe généralement au service informatique de l'établissement intéressé.

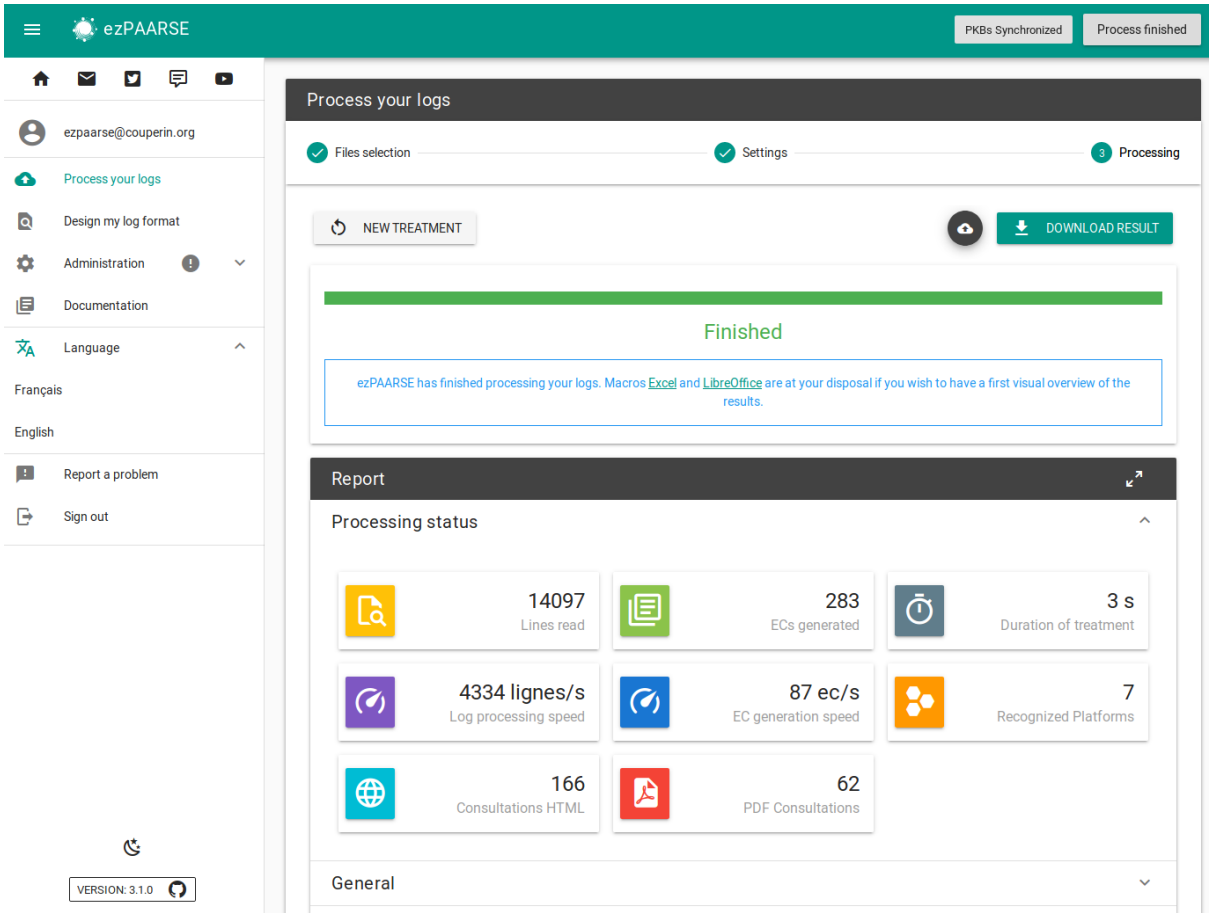

Figure 1 : un fichier logs correctement analysé par ezPAARSE

5 https://www.couperin.org/
6 https://www.inist.fr/qui/

2.2.2 Historique, difficultés, intérêt et solutions

Avant la création d'ezPAARSE, l'INIST-CNRS utilisait un système similaire, développé et utilisé en interne uniquement. La maintenance de cette application "legacy", très liée aux autres briques alors en place, représentait une charge de travail importante et difficilement soutenable pour l'unique personne qui y était associée. Ce système, jugé extrêmement intéressant par un grand nombre de collègues d'autres établissements, prouvait déjà que le mécanisme d'analyse des logs mis en œuvre était valide et opérationnel et pouvait donner des résultats d'une grande finesse, en particulier quand ils étaient injectés dans une solution flexible de visualisation, Omniscope Visokio à l'époque.

Partant de ce travail préalable, ezPAARSE est implémenté sous licence libre[7] et proposé gratuitement dès le début des développements. Techniquement repris à zéro avec des outils jugés plus adaptés[8], le projet est désormais mené en partenariat entre l'INIST-CNRS, l'Université de Lorraine et le consortium Couperin.org dans un mode de projet Agile qui permet, une fois la feuille de route générale (ou "vision du produit") arrêtée, de prendre en compte de nouvelles demandes de fonctionnalités au fil de l'eau et de les prioriser à nouveau à chaque itération[9].

La qualité des résultats fournis par ezPAARSE est basée sur un fort travail collaboratif pour étendre et tenir à jour ses capacités de reconnaissance des ressources consultées sur les différentes plateformes web des fournisseurs et éditeurs. Ce travail-là, identifié comme essentiel, est relativement découplé du cycle de développement d'ezPAARSE et pris en charge dans un environnement distinct, AnalogIST, décrit *infra*.

2.3 AnalogIST : un travail collaboratif et valorisé d'analyse des plateformes éditeurs

ezPAARSE est aujourd'hui livré avec une liste de plus de 200 parseurs, qui sont le résultat d'un travail d'analyse fourni par la communauté, et qui assurent aux établissements utilisateurs une bonne couverture de leurs abonnements aux ressources électroniques.

La plateforme communautaire, appelée AnalogIST[10], est l'endroit où les différents types d'URLs sont collectés, analysés semi-automatiquement et commentés pour que le parseur correspondant puisse être précis et complet [3]. Pour les éditeurs ou fournisseurs qui ne sont pas encore reconnus par ezPAARSE ou qui ont simplement besoin d'une mise à jour, il est facile de venir contribuer à une analyse de plateforme : pour cela, il suffit d'ouvrir un compte Trello et de suivre le guide[11] !

Une fois l'analyse jugée suffisante, ou complète, un informaticien (généralement de l'équipe ezPAARSE/ezMESURE) prend le relais pour implémenter le parseur correspondant. Ce parseur viendra s'ajouter aux parseurs existants dans le dépôt github dédié[12] et pourra être récupéré par toutes les instances d'ezPAARSE, par un mécanisme simple de mise à jour

7 Licence de type CeCiLL, équivalent à la GPL v3 en droit français : https://github.com/ezpaarse-project/ezpaarse/blob/master/LICENSE.txt

8 l'application d'origine était écrite en PHP. L'application actuelle utilise la stack Node.JS/Express/MongoDB

9 https://www.youtube.com/ezpaarse

10 http://analyses.ezpaarse.org

11 Comment commencer une analyse de plateforme sur AnalogIST (FAQ du blog) : https://blog.ezpaarse.org/2017/11/faq-comment-commencer-une-analyse-dans-la-plateforme-analogist/

12 https://github.com/ezpaarse-project/ezpaarse-platforms/

embarqué dans la zone d'administration du logiciel. Ce mode de fonctionnement garantit que chaque nouveau parseur, ainsi que chacune des mises à jour du logiciel et de ses ressources, profitent à toutes les installations existantes d'ezPAARSE : le travail n'est nécessaire qu'à un endroit et une seule fois.

Enfin, pour vérifier simplement que les parseurs livrés se comportent correctement, nous fournissons une extension de navigateur, ezlogger[13], qui capture le trafic et l'envoie vers une instance ezPAARSE (l'instance de démonstration[14] est choisie par défaut) qui traite en direct ces traces et génère des événements de consultation.

La participation de la communauté est encouragée et récompensée par des badges ouverts[15] qui en valident les différentes étapes : déclaration d'une nouvelle plateforme à prendre en compte, ajout d'une analyse, création ou mise à jour du parseur correspondant. Outre l'aspect ludique des badges, c'est une réelle compétence qui est identifiée dans les profils de poste des collaborateurs et utilisateurs de ces outils. Après un peu plus d'un an de fonctionnement, le serveur de badges[16] a déjà attribué 75 badges pour les analyses de plateformes et les créations de parseurs.

2.4 Le support et la formation

L'accompagnement et la formation sont une grande part de l'activité menée autour de la suite logicielle, qui se matérialise par différents supports accessibles aux utilisateurs :

- screencasts des étapes techniques (installation, utilisation, automatisation) sur YouTube,
- articles explicatifs dans le blog,
- séances de formations thématiques en présentiel ou par partage d'écran.

Des "rendez-vous ezPAARSE/ezMESURE" en ligne, d'une heure environ, sont régulièrement organisés pour les utilisateurs d'un établissement pour faire un point d'étape ou bénéficier d'une assistance à la mise en place de la chaîne de traitement : installation d'ezPAARSE sur un serveur local, automatisation des traitements, et comme il sera expliqué plus loin, automatisation des chargements dans ezMESURE et prise en main des tableaux de bord Kibana. A l'issue d'une séance, l'ensemble de la chaîne est généralement opérationnelle et l'établissement est en mesure d'avoir une vision des consultations de ses utilisateurs.

Le blog[17] rassemble une mine d'informations : des tutoriels techniques ou non, une rubrique FAQ, l'annonce des ateliers et les principales actions de communication. Toutes les mises à jour des plateformes y sont aussi consignées. Conçu en mode participatif modéré, la communauté peut l'enrichir et documenter leurs cas d'usage.

Enfin, la présence sur Twitter permet une communication directe avec la communauté pour répondre rapidement à des questions ponctuelles, pour rappeler des rendez-vous importants ou annoncer des nouvelles "fonctionnalités phares".

13 https://blog.ezpaarse.org/2018/08/tutoriels-tester-les-analyses-durl-dans-analogist-avec-ezlogger/
14 http://demo.ezpaarse.org
15 https://openbadges.org/
16 https://github.com/ezpaarse-project/ezpaarse-badge
17 https://blog.ezpaarse.org/

2.5 Enrichissement des données

La teneur en informations des fichiers logs pourrait paraître relativement pauvre de prime abord. S'y trouvent tout de même des éléments importants qui peuvent être utilisés pour enrichir davantage les événements de consultations :

- Les identifiants documentaires[18], propriétaires ou non, qu'utilise ezPAARSE pour interroger des référentiels comme celui de Crossref[19] et rapatrier des métadonnées bibliographiques explicites : année de publication, titre de la revue, domaine scientifique
- Le jeton d'authentification du lecteur attribué par l'établissement via son annuaire pour caractériser les cohortes d'utilisateurs en son sein : par statut, niveau d'étude, spécialité, etc.

ezPAARSE met en œuvre un mécanisme de mise en cache pour les données supplémentaires qu'il rapatrie et optimise les temps de traitement en économisant les appels multiples à ces sources de données externes.

2.6 La production de données n'est que la première étape

Pouvoir produire ce type de données n'est que la première étape, qui commence à être bien connue par les établissements partenaires, en particulier en France (notre cœur de cible historique) et à l'international, aux États-Unis en particulier où le logiciel a déjà été installé plus d'une centaine de fois et semble utilisé régulièrement.

La deuxième étape consiste maintenant à visualiser et analyser, d'une manière pratique et facile, les données produites par les différentes installation d'ezPAARSE. Et c'est là qu'intervient ezMESURE : référentiel national d'indexation et de stockage, on y trouve également un outil de visualisation pour l'assemblage de tableaux de bord dynamiques et personnalisables.

18 On pense en particulier au Title_ID tel que le décrit la norme KBART : https://www.uksg.org/kbart/s5/guidelines/data_fields

19 https://en.wikipedia.org/wiki/Crossref

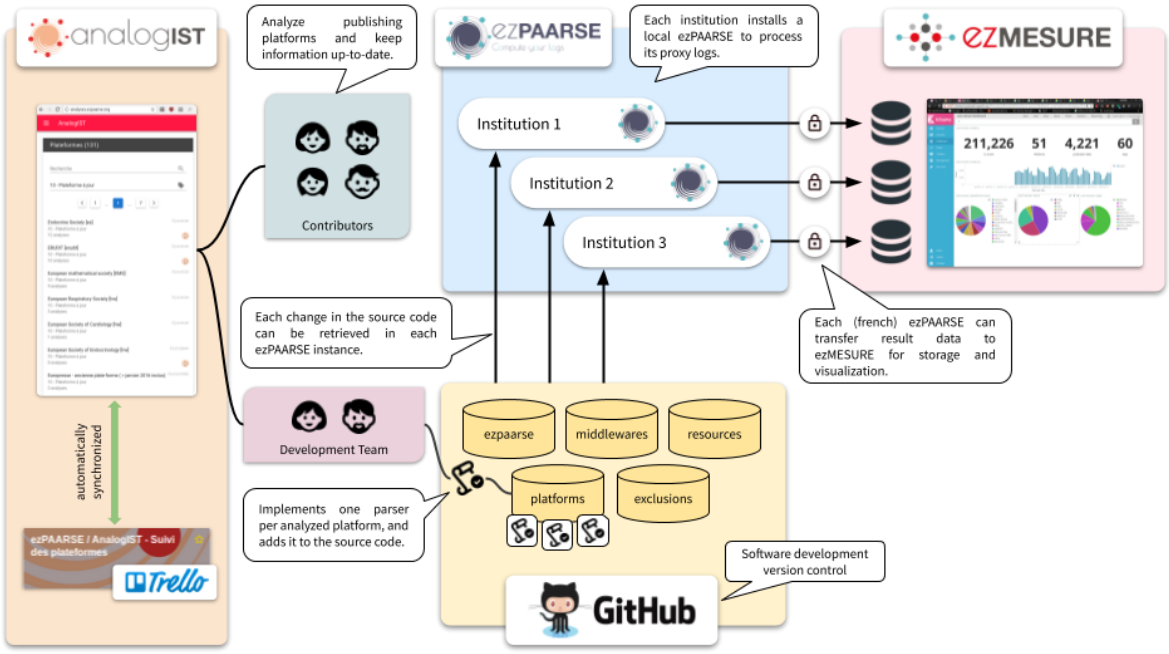

Figure 2 : vue d'ensemble du dispositif AnalogIST / ezPAARSE / ezMESURE et l'infrastructure sous-jacente Trello et GitHub

3 L'ENTREPOT NATIONAL EZMESURE

3.1 Présentation rapide

Depuis le début de l'année 2016, l'INIST-CNRS, Couperin.org et ses membres ont fourni un effort soutenu pour rassembler les événements de consultation générés et enrichis par les différentes installations d'ezPAARSE dans le référentiel national français ezMESURE[20] pour créer, consolider et représenter des tableaux de bord dynamiques. Le référentiel héberge maintenant les données d'accès de près de 60 institutions.

A l'heure actuelle, un établissement qui dépose ses données ezPAARSE dans ezMESURE est le seul à y accéder : son espace de stockage lui est réservé et les données qui s'affichent dans ses tableaux de bord ne sont visibles que par lui. C'est le principe qui s'applique par défaut.

Chaque institution désigne deux correspondants :

- un interlocuteur "technique" garant du bon fonctionnement de la suite logicielle en local
- un interlocuteur "documentaire" chargé de vérifier la couverture documentaire de son établissement et de gérer les accès aux données déposées dans ezMESURE pour les autres membres de son établissement.
  94 badges de correspondants ont été attribués, qui matérialisent là encore des activités spécifiques en lien avec le projet.

20 Documents descriptifs disponibles https://www.couperin.org/services-et-prospective/grilles-d-evaluation-ressources/134-statistiques-dusage/1274-ezmesure

Basé sur une suite ElasticSearch[21] et Kibana[22], ezMESURE est capable de stocker, récupérer et afficher les données d'usage dans une variété de tableaux de bord qui peuvent être adaptés aux besoins de chaque utilisateur et ouvre la voie à des vues agrégées à différents niveaux.

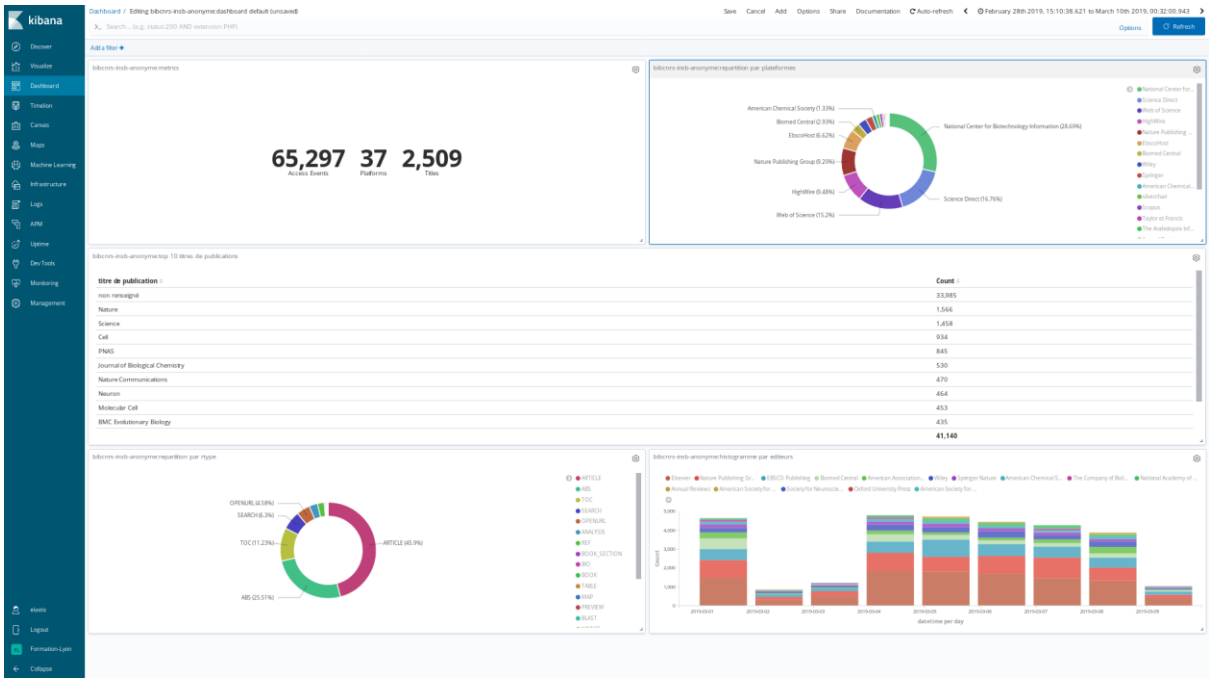

Figure 3 : Un tableau de bord Kibana dans ezMESURE

Avec cet écosystème, nous fournissons à nos utilisateurs une chaîne de traitement complète, pour analyser finement leurs données.

3.2 Les données collectées : tendre vers l'homogénéité

En fonction des établissements déposants, il peut y avoir de nettes différences dans la quantité et le niveau de finesse des données déposées : certains y ont déjà déposé plusieurs années de données, d'autres seulement quelques semaines ou quelques mois, pour tester le service. Certains ne sont abonnés qu'à une dizaine de ressources, d'autres à plus d'une centaine.

Des différences existent aussi dans la nature des données déposées : si les informations bibliographiques fournies par ezPAARSE sont rendues homogènes par le dispositif même, les données sur les profils des utilisateurs peuvent connaître de grandes variations. Certains établissements sont ainsi capables de rapatrier des données fines depuis leurs annuaires, quand d'autres ne fonctionnent que par grandes catégories (par exemple : "étudiant" ou "non étudiant"), ou parfois même sans cette information. Un travail de normalisation est en cours pour adopter une base commune minimale qui s'appuiera sur des recommandations en vigueur[23].

21 https://www.elastic.co/products/elasticsearch
22 https://www.elastic.co/products/kibana
23 par exemple, les recommandations SUPANN : https://services.renater.fr/documentation/supann/supann2018/recommandations2018/index

4 CHANTIERS EN COURS OU A VENIR

4.1 Tableaux de bord communs et partage de l'information

L'un des objectifs du service ezMESURE est de permettre la comparaison entre les établissements dans des tableaux de bord transverses. Que ce soit pour analyser l'usage d'une ressource particulière par tous ses abonnés, ou pour disposer d'informations sur tous les établissements d'une même zone géographique, ou pour conduire des études sur un champ disciplinaire, etc. L'outillage décrit ici rend la chose techniquement possible et les chantiers d'application ne manquent pas mais la quantité de données collectées est encore un peu limitée pour les mener à leur terme.

4.2 Reconnaître les ressources qui sont disponibles en Open Access

L'interrogation de référentiels externes (Crossref[24], Unpaywall[25], etc.) permet, dans de nombreux cas, de récupérer des informations sur le statut ou la licence Open Access des unités documentaires auxquelles accèdent les utilisateurs. Ces informations, rapatriées dans les données ezPAARSE, peuvent ensuite permettre d'observer sur ezMESURE la proportion d'articles disponibles en libre accès (ou sur le point de le devenir, dans le cas d'un embargo) sur une plateforme particulière ou un ensemble de plateformes, et d'observer son évolution dans le temps.

4.3 Mesurer le trafic sur les serveurs d'archives ouvertes

ezPAARSE fonctionne aussi avec des logs de serveurs d'archives ouvertes, sur le même principe. Un travail a été initié avec différentes acteurs français pour leur permettre de produire leurs propres données d'usage. Des problématiques spécifiques, liées à l'accès gratuit sont à traiter : le filtrage des robots en particulier, qu'il faut pouvoir distinguer des consultations "humaines", et l'impossibilité de pouvoir caractériser les lecteurs quand ils ne s'identifient pas ou que leur adresse IP ne correspond à aucune plage déclarée par les établissements.

5 UNE OFFRE DE SERVICE QUI MONTE EN PUISSANCE

Les solutions ezPAARSE et ezMESURE ont été conçues pour servir prioritairement les établissements de l'Enseignement Supérieur, de la Recherche et de l'Innovation en France mais constituent une offre de service généralisable pour :

- la mise à disposition d'un outil d'aide à la gestion de la politique documentaire
- la participation à un réseau de professionnels des données d'usage
- la valorisation des données d'usage et des investissements documentaires
- l'aide à la décision dans le cadre des politiques nationales de l'information scientifique et technique

Les chantiers d'application de ces composants, pour l'essentiel libres et gratuits, sont vastes.

24 https://github.com/CrossRef/rest-api-doc
25 https://unpaywall.org/products/api

References

[1] Bergstrom, T., Uhrig, R., & Antelman, K. (2018). Looking under the COUNTER for overcounted downloads. *UC Santa Barbara: Department of Economics*. Retrieved from https://escholarship.org/uc/item/0vf2k2p0

[2] Porquet, T., Lechaudel, D., Gully, S., & Jouneau, T. (2013). Le dispositif ezPAARSE/AnalogIST : pour une analyse mutualisée des logs d'accès aux ressources documentaires payantes. In JRES 2013. Retrieved from https://2013.jres.org/archives/33/paper33_article.pdf

[3] Lechaudel, D., Porquet, T., Fabry, C., Gully, S., Jouneau, T. & Schurter, Y. (2014) AnalogIST / ezPAARSE: analysing locally gathered logfiles to determine users' accesses to subscribed e-resources. In LIBER 43rd Annual Conference. Riga, Lettonie.