• SÉRIE DE RAPPORTS SUR LES DÉPÔTS OUVERTS

Statistiques des dépôts institutionnels : approches fiables et cohérentes pour les dépôts canadiens

par ***Will Roy, Brian Cameron et Tim Ribaric***
(membres du Sous-groupe sur les normes relatives aux données d'utilisation des dépôts institutionnels du Groupe de travail sur les dépôts ouverts de l'ABRC)

DÉCEMBRE 2019

www.carl-abrc.ca

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table des matières

Introduction

> « Les données d'utilisation sont un moyen efficace pour les bibliothèques de démontrer la valeur de leurs dépôts institutionnels. Toutefois, les outils actuels ne sont pas toujours fiables et peuvent donner lieu à une sous-estimation ou à une surévaluation du nombre de fichiers téléchargés. De plus, bien qu'il soit possible de consulter les statistiques au moyen de diverses interfaces de dépôt, sans norme reconnue, il est impossible d'évaluer et de comparer de façon fiable et significative les données d'utilisation entre les différents dépôts institutionnels[1]. »

Le Sous-groupe sur les normes relatives aux données d'utilisation des dépôts institutionnels du Groupe de travail sur les dépôts ouverts a lancé un exercice de rassemblement d'information pour mieux comprendre à la fois les pratiques existantes des dépôts canadiens et les nouveaux outils et processus dont disposent les dépôts pour un suivi et une surveillance plus efficaces. L'exercice est directement lié aux objectifs plus vastes du Groupe de travail sur les dépôts ouverts, soit « renforcer et ajouter de la valeur au réseau canadien de dépôts à libre accès en collaborant plus étroitement et en adoptant un éventail plus large de services[2]».

L'approche que nous recommandons est que tous les dépôts institutionnels canadiens adoptent collectivement les statistiques OpenAIRE. Elle s'harmonise avec les recommandations suivantes, que notre groupe propose par ailleurs :

Recommandations

Nous proposons les recommandations **obligatoires (O)** et **facultatives (F)** suivantes :

> ***R1(O) :*** *Tous les* dépôts institutionnels *canadiens devraient adopter le code de pratique COUNTER.*
>
> ***R2(O) :*** *Tous les* dépôts institutionnels *canadiens devraient choisir un service permettant l'interopérabilité avec d'autres services Web via une API entièrement ouverte, ou accessible, à base de permissions.*
>
> ***R3(O) :*** *Tous les* dépôts institutionnels *canadiens devraient utiliser un service de statistiques qui pratique la communication transparente et maintient une stratégie de gouvernance.*

---

[1] Mandat, sous-groupe sur les normes relatives aux données d'utilisation des dépôts institutionnels du groupe de travail sur les dépôts ouverts de l'ABRC, 2018. http://www.carl-abrc.ca/wp-content/uploads/2019/06/mandats_sousgroupes_GTDO_juin2019.pdf.

[2] Mandat, groupe de travail sur les dépôts ouverts de l'ABRC, 2018. http://www.carl-abrc.ca/wp-content/uploads/2018/03/ToR_ORWG_fre.pdf.

En outre, nous recommandons vivement que les dépôts institutionnels canadiens prennent en considération les conseils suivants :

> ***R4(F) :*** *Faire de nouveaux investissements pour comprendre et utiliser le format « NCSA Common log format » pour les fichiers journaux de serveur*
>
> ***R5(F) :*** *Effectuer des recherches sur les incidences en matière de confidentialité d'une collecte de statistiques d'utilisation via des services tiers avec des intérêts commerciaux et étudier les autres solutions possibles.*
>
> ***R6(F) :*** *Pratiquer un sain scepticisme à l'endroit des outils et des solutions qui promettent des statistiques d'utilisation « accrues » et préconiser plutôt une évaluation responsable des collections en considérant les aspects multiples de l'utilisation*

Le tableau suivant montre comment trois services existants ont été notés en fonction des caractéristiques souhaitées incluses dans les recommandations qui précèdent.

RECOMMANDATION	IRUS-UK	OpenAIRE	RAMP
N° 1 Conformité avec COUNTER	3	3	1
N° 2 Interopérabilité (c.-à-d. accès par API, données ouvertes, tableaux de bord, etc.)	3	3	2
N° 3 Gouvernance / transparence	3	3	3
N° 4 *utiliser le format « NCSA Common log format » pour les fichiers journaux de serveur*	S.O.	S.O.	S.O.
N° 5 Accent sur la protection de la vie privée	3	3	2
N° 6 Accent sur « l'exactitude » plutôt que sur la « quantité » des résultats	3	3	2
Catégorie spéciale : traitement des robots	3	3	2
Catégorie spéciale : facilité d'installation pour les dépôts institutionnels canadiens	2	2	3
Score total	**20**	**20**	**15**

Légende

3 = Approche favorable

2 = Approche acceptable

1 = À travailler

S.O. = Ne s'applique pas à la notation parce que la recommandation est pour les gestionnaires de dépôt plutôt que pour une fonction

On trouvera plus de contexte pour ces recommandations dans le rapport des constatations ci-dessous.

Rapport des constatations

> *« [traduction] En recueillant et présentant diverses mesures numériques de l'utilisation des dépôts, les gestionnaires de dépôt sont en mesure d'offrir un service précieux aux chercheurs et aux établissements[3]. »*

Contexte

Le présent rapport explore les statistiques d'utilisation en tant que mesure digne d'intérêt et de saisies par les gestionnaires de dépôt. Il vise d'autant plus à déterminer les moyens d'atteindre cet objectif. Les dépôts fonctionnent dans un écosystème savant diversifié avec un grand nombre d'intervenants; il est donc utile de tenir compte du contexte dans lequel les mesures peuvent être utiles à la grande collectivité des intervenants.

Définir et différencier les statistiques d'utilisation

Dans ce contexte, les statistiques d'utilisation mesurent les visionnements et les téléchargements d'un élément particulier hébergé soit sur une plateforme de publication, soit sur une autre plateforme de distribution numérique, comme un dépôt. Il existe deux méthodes pour saisir les statistiques d'utilisation : l'analyse des fichiers journaux, qui fait le suivi des événements d'utilisation du côté serveur, et le marquage de pages, pour le suivi des événements d'utilisation du côté client[4]. Les statistiques d'utilisation reflètent l'utilisation dont le point d'origine est un point d'accès particulier, bien que d'autres efforts tentent d'afficher et de combiner des statistiques d'utilisation à de multiples points d'accès, comme par exemple le projet Distributed Usage Logging [marquage d'utilisation distribué][5] par COUNTER et CrossRef et l'outil Lagotto du projet de la Public Library of Science[6].

Les statistiques d'utilisation reflètent une couverture des utilisateurs qui est différente des indicateurs fondés sur les citations. Au contraire des mesures par les citations, qui sont fondées sur les activités des auteurs, « [traduction] de nombreux utilisateurs

---

[3] Confederation of Open Access Repositories (COAR). 2013. Incentives, integration, and mediation: Sustainable practices for populating repositories. Repéré à : https://www.coar-repositories.org/files/Sustainable-best-practices_final2.pdf.

[4] COUNTER. 2019. The COUNTER code of practice for release 5. Repéré à : https://www.projectcounter.org/code-of-practice-five-sections/6-logging-usage/.

[5] https://www.crossref.org/working-groups/distributed-usage-logging/.

[6] http://www.lagotto.io/.

possibles (étudiants, décideurs, public intéressé) qui lisent les publications ou utilisent les données sans jamais publier. En outre, ce qu'une chercheuse peut lire ne se retrouve pas forcément dans ses publications »[7]. La Public Library of Science a démontré les limites des mesures fondées sur les citations en disant que « [traduction] seulement un utilisateur sur 70 qui télécharge un PDF en tirera une citation »[8]. Considérant les nombreux types d'utilisateurs possibles qui pourraient lire et consulter des articles sans avoir l'intention de les citer dans un texte de recherche, les statistiques d'utilisation offrent un nouveau point de vue prometteur pour juger de l'incidence du libre accès.

Qui en profite? Valeurs pour les intervenants

Il y a au moins huit groupes d'intervenants primaires susceptibles de profiter des mesures numériques émergentes. La National Information Standards Organization (NISO) a réalisé un projet des mesures de rechange de l'évaluation en 2016[9] et mis sur pied plusieurs groupes de travail visant à explorer la valeur de nouvelles mesures possibles où sont définis les huit groupes d'intervenants suivants : bibliothécaires, administrateurs de la recherche, comités de recrutement, organismes de financement, chercheurs universitaires, maisons d'édition/éditeurs, agents des médias, et fournisseurs de plateforme. Ils ont formulé des cas d'utilisation pour expliquer comment chaque groupe d'intervenants pouvait profiter des mesures proposées, puis les ont subdivisés en trois grands thèmes, à savoir : présentation des réalisations, réalisation des évaluations de la recherche et amélioration des découvertes[10].

Notre groupe a également constaté que les statistiques d'utilisation ont été présentées comme puissant outil de suivi et de compréhension des investissements, comme moyen de promouvoir l'avancement de la science ouverte, et comme outil qui s'inscrit dans le besoin exprimé par les conseils subventionnaires de la recherche de nouvelles mesures ouvertes, interopérables et à base de normes.

Statistiques d'utilisation pour aider à comprendre les investissements

Les universités et les organismes de financement sont toujours en quête de nouvelles méthodes pour établir le rendement du capital investi. Toute mesure chiffrée qui peut

---

[7] European Commission (2017). Next-generation metrics: Responsible metrics and evaluation for open science. Repéré à : https://ec.europa.eu/research/openscience/pdf/report.pdf.

[8] Lin, J. & Fenner, M. (2013). Altmetrics in evolution: Defining and redefining the ontology of article-level metrics. *Information Standards Quarterly*, 25(2). Repéré à : https://www.niso.org/sites/default/files/stories/2017-08/IP_Lin_Fenner_PLOS_altmetrics_isqv25no2.pdf.

[9] National Information Standards Organization (2016). Outputs of the NISO alternative assessment metrics project. NISO RP-25-2016. Repéré à : https://www.niso.org/publications/rp-25-2016-altmetrics.

[10] Ibid.

contribuer à faire comprendre le rendement, aux niveaux individuel et institutionnel, est utile, surtout si elle va au-delà des mesures traditionnelles des revues[11].

Dans le contexte de la science ouverte, on observe l'émergence d'un nouveau besoin de mesures qui s'appliqueraient non seulement à l'*offre* dans le contexte de la recherche, mais aussi à la *demande*[12]. Les statistiques d'utilisation représentent une *mesure fondée sur le lecteur* plutôt qu'une mesure fondée sur l'auteur, comme le facteur des citations, si bien qu'il s'en dégage des perceptions différentes de l'incidence de la recherche et des voies d'analyse divergentes auxquelles se prêtent les outils de mesure traditionnels. Par exemple, les bibliothèques ont toujours trouvé que les statistiques d'utilisation étaient très utiles pour éclairer les décisions d'acquisition[13], ce qui pourrait s'expliquer en partie par leur capacité de démontrer l'utilisation et la demande propres à leur collectivité locale.

La demande de produits de recherche à financement ouvert peut être aussi démontrée par des statistiques d'utilisation qui sont normalisées et agrégées à l'échelle de nombreux fournisseurs de dépôts ouverts.

Statistiques d'utilisation pour promouvoir l'ouverture

Dans le cas du rôle joué par les mesures numériques dans le soutien et la stimulation de la science ouverte, la Commission européenne a aussi estimé que les mesures numériques peuvent servir deux objectifs principaux dans la promotion du soutien de la science ouverte[14]. Ces objectifs sont les suivants :

- surveiller le développement du système scientifique vers l'ouverture à tous les niveaux;
- mesurer le rendement afin de récompenser l'amélioration des façons de travailler au niveau des groupes et des personnes[15].

Pour atteindre ces objectifs, il est fortement recommandé de créer de nouveaux indicateurs et d'en faire une utilisation responsable, dans le sens des explications

---

[11] Organ, M. (2006). Download statistics - what do they tell us?: The example of research online, the open access institutional repository at the University of Wollongong, Australia. *D-Lib Magazine*, 12(11) https://doi.org/10.1045/november2006-organ.

[12] European Commission ... Ibid.

[13] Glänzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: Confusing terminology? *Scientometrics*, *102*(3), 2161–2164. https://doi.org/10.1007/s11192-014-1472-7.

[14] European Commission... Ibid.

[15] Dans ce contexte, la mesure du rendement ne s'applique pas aux résultats des chercheurs, mais plutôt aux façons de mesurer les pratiques ouvertes qui ne sont pas prises en compte ni reconnues dans les structures de récompenses traditionnelles.

données dans des documents comme le Manifeste de Leiden[16], le rapport Metric Tide[17] et la Déclaration de San Francisco sur l'évaluation de la recherche[18]. Tous ces rapports et initiatives influents préconisent l'élaboration et le déploiement de normes, la transparence et l'ouverture, l'interopérabilité et l'utilisation responsable des mesures métriques.

Il y a aussi des incitations pour les auteurs qui peuvent encourager une plus grande participation à l'ouverture. Le rapport final de l'analyse des statistiques d'utilisation du JISC plaide que la valeur des statistiques d'utilisation des dépôts réside dans leur actualité. Alors qu'il faut du temps pour agréger les mesures traditionnelles applicables aux citations, les données d'utilisation des dépôts sont présentées plus ou moins instantanément, ce qui permet aux auteurs d'évaluer plus immédiatement la visibilité de leurs travaux[19]. Comme telle, la fonction des mesures applicables aux dépôts est un outil de recrutement. Bruns et Inefuku (2015) soutiennent que « [traduction] la collecte et la production de données de mesure numériques sont des outils précieux que les gestionnaires de dépôt peuvent exploiter pour soutenir et encourager la participation du corps professoral aux dépôts »[20].

Le besoin de mesures fondées sur des normes ouvertes et interopérables

Il y a une foule d'autres mesures possibles à prendre en compte, sans compter le besoin de prendre des décisions éclairées sur ce qu'il faut faire pour s'assurer que ces mesures sont bien générées, organisées et utilisées. En 2015, le Higher Education Funding Council of England a commandé un examen indépendant du rôle des mesures numériques dans l'évaluation et la gestion de la recherche. Il résume la tendance en ces termes : « [traduction] Il y a de puissants courants qui font monter la marée des mesures numériques. Qu'il suffise de nommer les pressions croissantes pour l'audit et l'évaluation des dépenses publiques consacrées à l'enseignement et à la recherche. » Le rapport ajoute que les administrateurs de la recherche expriment clairement désormais le besoin d'indicateurs « reposant sur une infrastructure de données ouverte et interopérable »[21].

---

[16] Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature News, 520*(7548), 429. https://doi.org/10.1038/520429a.

[17] Wilsdon, J., L. Allen, E. Belfiore, & R. Kain. (2015). The Metric tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. https://doi.org/10.13140/RG.2.1.4929.1363.

[18] DORA (2012). San Francisco Declaration on Research Assessment. Repéré à : https://sfdora.org/read/.

[19] Joint Information Systems Committee (2008). Final Report: JISC Usage Statistics Review. Repéré à : https://repository.jisc.ac.uk/250/1/Usage_Statistics_Review_Final_report.pdf.

[20] Bruns, T. & Inefuku, H. W. (2015). Purposeful metrics: Matching institutional repository metrics to purpose and audience. *Digital Scholarship and Initiatives Publications, 4*. Repéré à : https://lib.dr.iastate.edu/digirep_pubs/4.

[21] Wilsdon, J., Allen, L., Belfiore, E., & Kain, R. (2015)... Ibid.

Il y a une excellente occasion de faire participer les dépôts à cette *marée de mesures numériques* et nous pouvons faire en sorte qu'elles soient utiles et qu'elles respectent les normes d'assurance de la qualité.

Les justifications qui précèdent pour la collecte de données d'utilisation sont plus faciles à expliquer et plus utiles si les statistiques sont comparables et fiables. La comparaison avec les éditeurs est alors faisable et rentable[22].

Les statistiques d'utilisation recueillies, et présentées dans nos dépôts, ne répondent tout simplement pas à cette norme à l'heure actuelle. Les nouveaux outils et les nouvelles approches explorés dans le présent rapport nous aideront à franchir les prochaines étapes nécessaires pour répondre à ces exigences.

Processus

Au printemps 2018, nous avons entrepris deux volets d'activités.

La première activité a consisté à élaborer et distribuer un questionnaire s'adressant aux gestionnaires des dépôts canadiens. L'objectif visé par la distribution de ce questionnaire d'enquête était d'aider notre groupe à comprendre les pratiques existantes de collecte et de mesurer les statistiques d'utilisation en contexte canadien, et à mieux comprendre les buts et les aspirations exprimés qui sont associés à la collecte des statistiques d'utilisation. Les conclusions de cette activité sont présentées dans la section Réponses à l'enquête sur les statistiques d'utilisation des dépôts institutionnels.

Dans la deuxième activité, nous avons analysé les outils émergents, comme IRUS-UK, les statistiques OpenAIRE et le Repository Analytics & Metrics Portal (RAMP). En parallèle, nous nous sommes penchés sur la version 5 du code de pratique COUNTER, qui comprend des pratiques et des recommandations pour le blocage des robots, la suppression des données d'utilisation inexactes, et la présentation des statistiques d'utilisation dans un format fiable et comparable. La section Normes et approches des données d'utilisation présente brièvement les services et les concepts que nous avons analysés.

Pour formuler ces recommandations, nous avons exploré les théories sous-jacentes, les pratiques exemplaires et les enjeux connus qui accompagnent la collecte des statistiques d'utilisation pour les dépôts. Les perceptions dégagées de ce processus sont reflétées dans le présent rapport et actualisées dans nos recommandations.

---

[22] MacIntyre, R. & Jones, H. (2016) IRUS-UK: Improving understanding of the value and impact of institutional repositories, *The Serials Librarian*, *70*(1-4), 100-105, https://doi.org/10.1080/0361526X.2016.1148423.

Réponses à l'enquête sur les statistiques d'utilisation des DI

Méthode

Une enquête par sondage a fait comprendre le paysage et la pratique actuelle des établissements de l'ABRC en ce qui concerne les statistiques et les analyses des dépôts. On trouvera le questionnaire de l'enquête à l'annexe du présent document. Le questionnaire, qui comporte dix questions, a été présenté dans les deux langues officielles. Chacun des établissements de l'ABRC a été invité à donner une seule réponse. Au moment de la préparation des statistiques ci-après, 42 réponses avaient été reçues, ce qui donne un taux de réponse de 70 %. En général, les réponses pourraient être classées par thème.

Sommaire

Modes de collecte et d'utilisation en général

La figure 1 présente un sommaire des résultats. En général, nous avons observé que :

- 2,5 % des répondants (n=1) ne faisaient pas de suivi de l'utilisation des dépôts.
- 75 % (n=30) des répondants partageaient des statistiques avec l'ensemble de la collectivité du campus.
- 47,5 % (n=19) des répondants ont recueilli et utilisé les fichiers journaux comme source de statistiques d'utilisation.

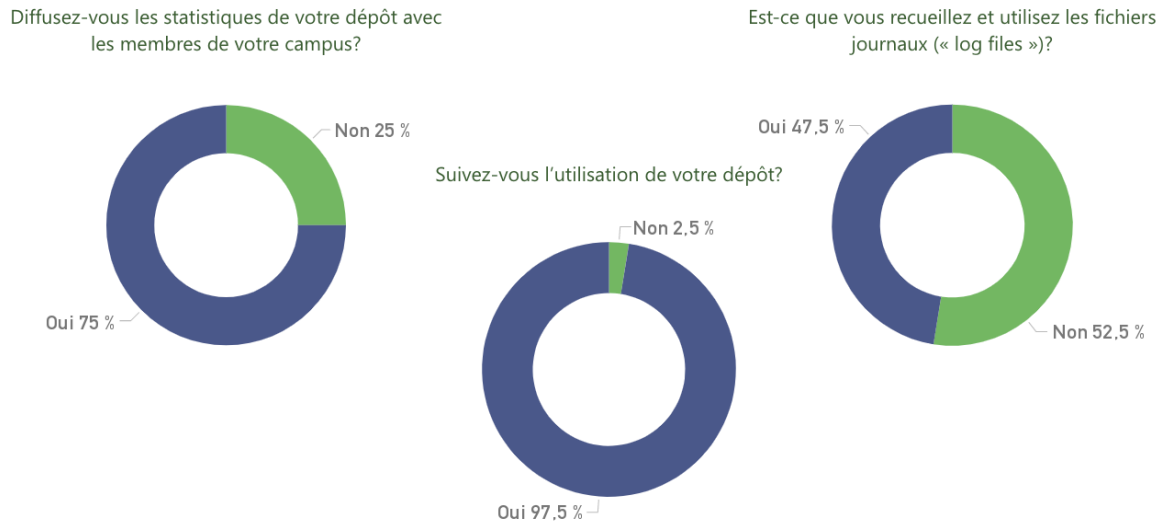

Figure 1
Collecte et partage de données dans les dépôts institutionnels canadiens

Utilisation des services de Google

Google offre gratuitement une vaste suite d'outils et de méthodes disponibles, que les créateurs de contenu Web peuvent utiliser pour effectuer une optimisation modeste des moteurs de recherche et une analyse de l'utilisation. La figure 2 résume l'utilisation

des services de Google par les répondants. Sans grande surprise, la majorité des établissements, 85 % (n=34) utilisent Google Analytics pour produire des rapports permettant d'accéder à de l'information sur l'utilisation, en temps réel. Il y a deux autres méthodes, moins populaires : l'envoi à Google des plans de sites, dont profitent seulement 37,5 % (n=15), et la mise en œuvre de la console Google Search, utilisée par 45 % (n=18). Le plus faible taux d'utilisation de ces deux derniers services pourrait signifier que les administrateurs de dépôt ne les aiment pas ou ne les connaissent pas. Puisque le coût n'est pas un facteur vu que tous ces services sont gratuits, il est raisonnable de penser que ces services pourraient être utilisés plus largement s'ils faisaient l'objet d'une meilleure promotion auprès des administrateurs de dépôt.

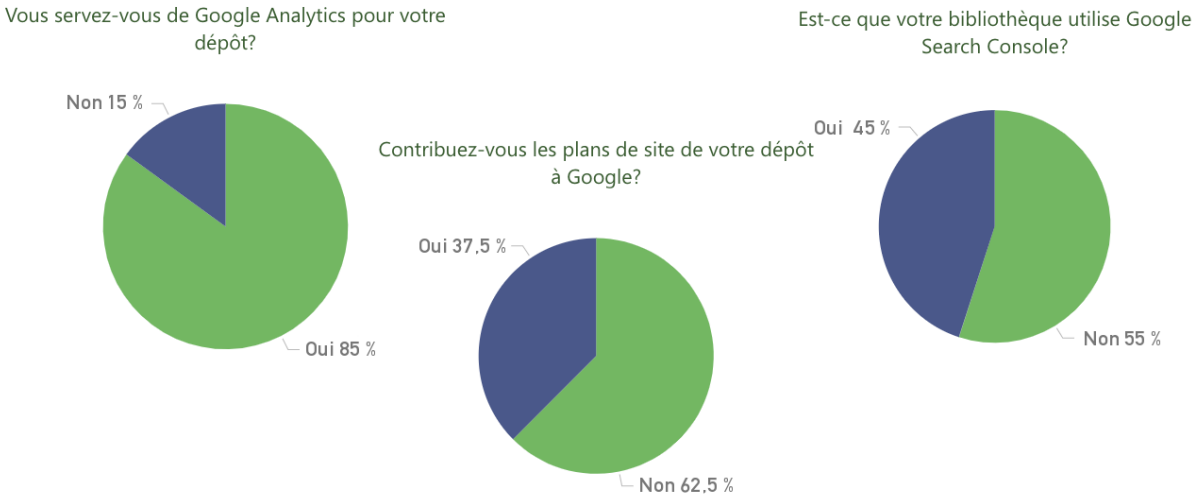

Figure 2
Utilisation de services Google dans les dépôts institutionnels

Méthodes d'utilisation et sources de renseignements sur l'utilisation

Les répondants ont été également invités à décrire les diverses méthodes de saisie des statistiques d'utilisation et à préciser si le trafic automatisé des robots était bloqué. La figure 3 résume les réponses. Les analyses classées comme internes (comprises dans le dépôt) viennent des systèmes d'analyse intégrés fournis dans la plateforme du dépôt. Les systèmes externes sont les systèmes d'analyse fournis par Google, ou par une analyse secondaire des fichiers journaux, ou grâce à d'autres produits statistiques décrits dans le présent rapport. 52 % (n=21) des dépôts font appel à des outils d'analyse internes et externes. Le filtrage des robots (trafic automatisé) fait ressortir une répartition presque égale. Cette source possible de bruit n'est éliminée entièrement que pour 47,5 % (n=19) des cas.

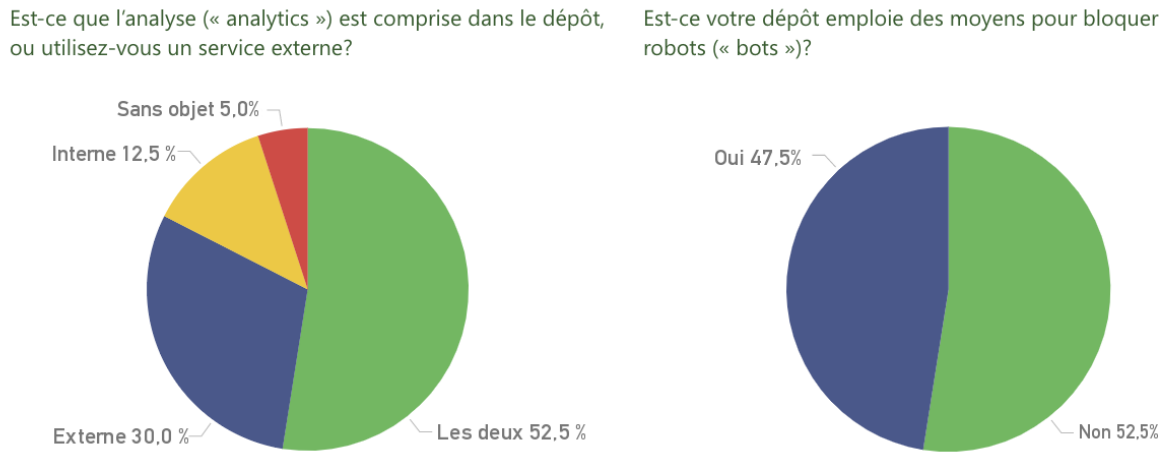

Figure 3
Aspects techniques des dépôts institutionnels

Satisfaction à l'égard des renseignements fournis

Les répondants ont été invités à commenter l'utilité perçue des statistiques d'utilisation qui ont été partagées avec les intervenants. Les réponses ont varié, mais la majorité des répondants étaient au moins satisfaits des services de statistiques des dépôts. Cependant, un quart des répondants n'ont pas répondu, ce qui pourrait indiquer qu'ils n'avaient pas suffisamment d'expérience de cet aspect de la consultation des intervenants pour se permettre des commentaires. La figure 4 présente le ratio des réponses à cette question. 57,5 % (n=23) des réponses ont indiqué un niveau de satisfaction minimal chez les utilisateurs externes à l'égard des statistiques d'utilisation fournies, ce qui indique clairement qu'on peut faire mieux de ces services.

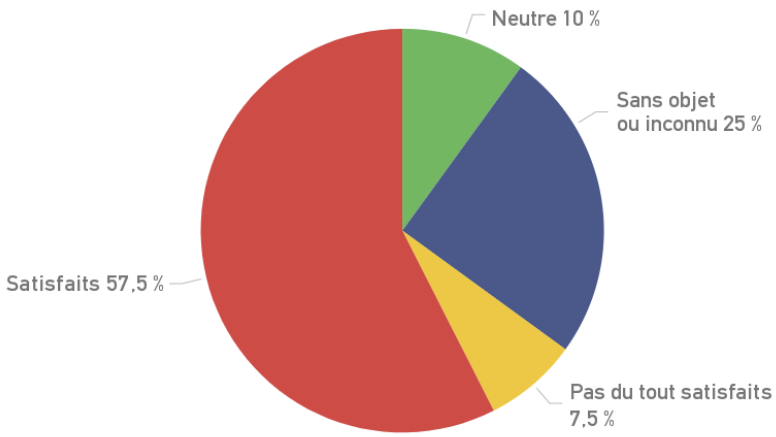

Figure 4

Quelle est votre perception de la satisfaction de vos utilisateurs par rapport aux statistiques que vous leur fournissez?

Normes et approches relatives aux données d'utilisation des DI

RAMP

Le Repository Analytics & Metrics Portal (RAMP) est un service Web issu d'un partenariat entre la Montana State University, l'Association of Research Libraries, l'Université du Nouveau-Mexique et OCLC Research.

L'équipe RAMP a généré un important corpus de recherche qui explique et confirme la sous-estimation présumée des événements de téléchargement lors de l'utilisation de Google Analytics, dont on sait qu'il ne compte que les événements de téléchargement déclenchés par une activité découlant directement du site Web d'un dépôt. Le RAMP compte les renvois externes (téléchargements directs) provenant de recherches sur Google en incorporant les données recueillies à partir d'un autre produit Google appelé Google Search Console (GSC). Les données de GSC combinées avec les données de Google Analytics et corrigées des doubles offrent un compte d'utilisation plus précis et probablement plus élevé des articles.

L'approche RAMP établit une distinction significative entre les pages, les pages sommaires et les téléchargements de contenu citables (TCC). Le RAMP présente le TCC comme l'indicateur le plus précieux de l'utilisation savante, parce qu'il est associé à l'accès direct à l'objet de recherche.

Le RAMP n'emploie pas ses propres tactiques pour bloquer les robots, préférant considérer Google comme « l'un des meilleurs au monde en détection des robots », et justifie son choix par la présomption selon laquelle « les conditions du marché donnent à Google des incitations et des ressources pour investir dans la détection des robots qui dépassent de loin les capacités des bibliothèques »[23].

IRUS UK (Institutional Repository Usage Statistics UK)

IRUS UK est un service offert aux membres du JISC (Joint Information System Committee) dans le cadre d'un abonnement au JISC. IRUS UK est financé par le JISC et il est actuellement utilisé dans au moins 144 établissements[24].

IRUS UK fournit des statistiques d'utilisation conformes à COUNTER pour les documents téléchargés des dépôts institutionnels participants via un plugiciel de code de suivi, qui sera bientôt révisé en fonction de la version 5 de COUNTER. L'un des principaux avantages de ce service est qu'il fournit des mesures comparables,

---

[23] OBrien, P., Arlitsch, K. Mixter, J., Wheeler, J. & Sterman, L. (2017) RAMP – the Repository Analytics and Metrics Portal: A prototype web service that accurately counts item downloads from institutional repositories, *Library Hi Tech, 35*(1), 144-158, https://doi.org/10.1108/LHT-11-2016-0122.

[24] Selon OpenAIRE (n.d.), National Open Access Desk, Source: https://www.openaire.eu/item/united-kingdom.

convergentes, complètes et à base de normes qui aident à déterminer la valeur et l'incidence des DI[25].

Les établissements de l'extérieur du Royaume-Uni peuvent utiliser IRUS UK comme service hébergé, moyennant frais. Les données sont transmises à OpenAIRE pour tout dépôt utilisant IRUS, ce qui dispense les DI de devoir gérer la conformité à OpenAIRE. L'intégration à OpenAIRE serait possible pour tout établissement canadien collaborant avec IRUS UK. L'étude de cas « *International collaboration and value: working with OpenAIRE* » fournit plus de détails sur l'intégration OpenAIRE[26].

IRUS UK est maintenant ouvert aux établissements de l'extérieur du Royaume-Uni et un projet pilote américain a eu lieu en 2018. L'OAPEN, l'Université d'Amsterdam et divers dépôts australiens et américains utilisent actuellement ce service. IRUS étudie aujourd'hui d'autres options concernant la prestation du service en Amérique du Nord.

Des rustines sont actuellement disponibles pour DSpace. Les autres options comprennent des plugiciels pour EPrints et Haplo, et un Ruby Gem pour les dépôts Hydra et Samvera. Pure et Worktribe ont mis en œuvre une fonctionnalité de suivi pour leurs plateformes logicielles de dépôt. IRUS a également eu des conversations avec d'autres fournisseurs, comme Bepress[27].

En date de 2018, Hyrax ne comprend pas d'interface utilisateur de l'administrateur, pour la restauration d'objets à partir de sauvegardes si l'on détecte des bitrots ou de la corruption de fichiers[28].

Statistiques OpenAIRE

OpenAIRE est une organisation européenne qui se voue à l'évolution de la communication savante vers l'ouverture par diverses initiatives différentes, qui comprennent la création de politiques, l'aménagement d'infrastructures, la formation, la défense et la promotion des intérêts, l'élaboration de normes et, plus particulièrement, les services pour l'interopérabilité des dépôts et l'utilisation des dépôts[29].

---

[25] IRUS-UK (n.d.). Frequently asked questions. Repéré à : http://irus.mimas.ac.uk/support/faqs/.

[26] IRUS-UK (2018). International collaboration and value: working with OpenAIRE case study. Repéré à : https://irus.jisc.ac.uk/documents/IRUS-UK_working_with_OpenAIRE.pdf.

[27] Voir : http://irus.mimas.ac.uk/ pour les guides, la boîte à outils, le soutien, les études de cas, les cas d'utilisation, les FAQ et les conseils.

[28] Rochkind, J. (2017). Exploring and planning with Sufia/Hyrax/Fedora fixity validation. https://bibwild.wordpress.com/2017/05/01/exploring-and-planning-with-sufiahyraxfedora-fixity-validation/.

[29] OpenAIRE (n.d.). Mission and Vision. (site Web). Repéré à : https://www.openaire.eu/mission-and-vision.

Plus particulièrement, OpenAIRE offre le service de statistiques d'utilisation OpenAIRE[30], un tableau de bord complet d'analyse de l'utilisation des dépôts sur la plateforme d'analyse Web de Matomo[31].

Les statistiques elles-mêmes sont produites selon les directives du code de pratique COUNTER. Comme la plupart des offres OpenAIRE, le service de statistiques d'utilisation est très complet. Par exemple, il peut effectuer l'élimination des doubles d'un même élément dans des dépôts multiples pour permettre d'agréger les rapports et de les fusionner.

Project COUNTER

Project COUNTER, fondé en 2003, est une organisation sans but lucratif internationale qui met au point une norme de code de pratique pour la production de données d'utilisation. Cette norme est un protocole bien connu et fiable, en usage dans de nombreuses bibliothèques et chez de nombreux fournisseurs à l'échelle internationale. En utilisant le protocole Z39.930-2014 de l'ANSI/NISO, il compile les données d'utilisation des ressources électroniques suivantes dans les bibliothèques : revues, bases de données, ensembles de données, livres, segments de livres, ouvrages de référence, bases de données multimédias, journaux, objets de dépôt, rapports et thèses ou dissertations.

La norme vise à fournir des « données d'utilisation convergentes, crédibles et comparables »[32]. Ce service aide les bibliothécaires à établir la valeur des ressources et les éditeurs à contribuer à cet objectif en fournissant des statistiques comparables entre les divers fournisseurs des bibliothèques. Le processus est soutenu par deux auditeurs approuvés de COUNTER. La conformité à la version 5, qui vise à réduire la complexité du code, à répondre aux besoins changeants, à assurer une plus grande personnalisation et à simplifier la maintenance, était requise pour janvier 2019.

En plus des statistiques d'utilisation, le cadre du code de pratique COUNTER définit les éléments de données à mesurer, les définitions de ces éléments, les rapports d'utilisation, les spécifications pour le traitement des données, les exigences pour le processus d'audit, et les lignes directrices pour l'élimination des doubles comptes.

COUNTER est complété par l'Initiative de récolte uniformisée des statistiques d'utilisation – SUSHI (Standardized Usage Statistics Harvesting Initiative), qui facilite l'extraction des données d'utilisation COUNTER, éliminant de ce fait le besoin d'extraire séparément les données de chaque site Web, et l'outil de validation des

---

[30] http://catalogue.openaire.eu/service/openaire.openaire_usage_statistics.

[31] https://matomo.org/.

[32] COUNTER (n.d.) Page de renvoi. (site Web). https://www.projectcounter.org/.

rapports COUNTER, qui permet aux fournisseurs et aux bibliothèques de tester la mise en œuvre des rapports SUSHI et COUNTER.

Certains services statistiques passés en revue dans le présent document sont fondés sur le protocole COUNTER, qui confère un haut degré de confiance dans la méthodologie de collecte des données et de production de rapports COUNTER.

Google Analytics

Google Analytics, lancé à la fin de 2005, est un service d'analyse Web qui recueille, mesure et analyse le trafic Web au moyen d'un code de suivi. En plus d'aider à comprendre le rendement de l'investissement, il est un outil d'étude de marché et un moyen d'optimiser l'efficacité des sites Web.

Google Analytics se concentre sur les dimensions et les mesures numériques. Les dimensions présentent les attributs des données, comme l'origine géographique du trafic et la page mesurée. Les mesures numériques sont des mesures quantitatives de ces données, comme le taux de rebond, la durée de la session, les pages par session, le temps moyen par page, le pourcentage de sortie, etc. Les données sont classées comme suit : acquisition, c'est-à-dire comment relever le trafic du site Web; le comportement, c'est-à-dire ce que les visiteurs font sur le site Web; et les conversions, c'est-à-dire une activité terminée.

Plusieurs problèmes peuvent entraver la collecte des données. Les plus fréquents sont les navigateurs avec JavaScript désactivé, les utilisateurs qui refusent les témoins, l'utilisation de deux appareils différents par le même utilisateur, ainsi que le filtrage de la publicité, ou l'utilisation de réseaux privés, qui empêchent dans les deux cas la collecte de certaines données. Google Analytics ne présente pas de données en temps réel et peut inclure des pourriels de renvoi. En outre, Google peut utiliser l'échantillonnage de données pour des sites Web très actifs.

Selon la recherche, Google Analytics ne convient pas aux dépôts institutionnels parce qu'il « ne saisit pas la vaste majorité des téléchargements de contenu citable non HTML » à partir des dépôts. Malgré cela, l'outil est utilisé dans un grand nombre de dépôts[33].

Analyse des fichiers journaux

Une caractéristique clé des systèmes qui fonctionnent comme serveurs Web est qu'ils génèrent un fichier contenant toutes les transactions qu'ils effectuent. Ce fichier est rédigé en clair et s'appelle un fichier journal. Il est structuré selon un format

---

[33] Obrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J., & Borda, S. (2016). Undercounting file downloads from institutional repositories. *Journal of Library Administration, 56*(7), 854-874. https://doi.org/10.1080/01930826.2016.1216224.

normalisé[34] appelé le format registre commun (CLF). Chaque ligne d'un fichier CLF est délimitée et formée des éléments essentiels de chaque action exécutée par le serveur. Ces composantes comprennent l'adresse IP, le timbre de la date et de l'heure, le fichier servi dans la demande, le code d'état, le volume du fichier en octets, et ainsi de suite.

Puisqu'ils décrivent chaque action qu'accomplit le serveur, ces fichiers journaux deviennent souvent très volumineux et peuvent rapidement atteindre le niveau des gigaoctets. À cause de leur verbosité possible, l'examen direct est fastidieux. En réponse à ce problème, de nombreux progiciels ont été créés pour l'agrégation et la présentation des tendances relevées dans ces fichiers. Les plateformes populaires comprennent : Analog[35] et Graylog[36]. Une bonne proportion de ces plateformes sont gratuites et à source ouverte, mais il existe aussi de nombreuses solutions commerciales. Les types d'analyse qu'elles peuvent fournir augmentent, mais comprennent des choses comme l'emplacement géographique des visiteurs (selon l'adresse IP), l'indication du contenu populaire, et la fourniture de renseignements sur les erreurs susceptibles de se produire, comme lorsque les utilisateurs demandent du contenu qui n'existe pas (p. ex., erreurs 404).

Vu qu'ils sont essentiellement des serveurs Web pour un type particulier de contenu, les dépôts génèrent des fichiers CLF qui peuvent contenir des renseignements sur l'utilisation. Par contre, il est difficile d'accéder aux fichiers mêmes, ce qui pourrait se révéler préjudiciable. Par exemple, avec une plateforme de dépôt en nuage, le fournisseur d'hébergement pourrait ne pas avoir de mécanisme pour exposer les fichiers journaux pour analyse. En outre, l'exécution du logiciel qui effectue l'analyse impose un fardeau supplémentaire à la Bibliothèque, qui doit désormais faire tourner une autre plateforme.

ROBOTS (bots)

La tendance à l'enclenchement du trafic sur le Web par des programmes informatiques, aussi appelés robots, peut rendre difficile la distinction ou la différenciation de l'utilisation humaine réelle, et ainsi créer de l'incertitude quant à l'exactitude des statistiques fournies. « Les robots sont plus nombreux que les

---

[34] IBM. Log file formats: NCSA Common. http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html#common.

[35] Analog CE. https://www.c-amie.co.uk/software/analog/.

[36] Graylog. https://www.graylog.org/.

humains selon un ratio de 10:1 au niveau des sessions, de 5:4 au niveau des accès HTTP bruts, et de 4:1 au niveau des mégaoctets transférés »[37].

La solution acceptée dans la plupart des services Web est d'obliger à utiliser un nom d'utilisateur et un mot de passe pour accéder au contenu[38]. Les collections ouvertes qui se veulent libres d'obstacles à l'accès n'ont pas cette option, et sont plutôt confrontées à la lourde tâche de bloquer les robots proactivement ou de filtrer les résultats pour éliminer rétroactivement le trafic des robots.

COUNTER fournit une liste de robots bien connus, dont il y aurait lieu de supprimer l'utilisation pour satisfaire aux exigences du processus d'audit. La liste « ne sera pas une liste exhaustive. Le besoin de règles et de processus plus complexes est bien compris »[39]. La collectivité tient aussi des listes de robots et d'araignées connus qui peuvent s'ajouter à un fichier sur le serveur Web, sous le nom robots.txt. En outre, la collectivité tient des listes d'adresses IP malveillantes par le projet Honeypot, qui peuvent servir à entraîner et à construire de meilleurs filtres.

Pour les services qui utilisent Google Analytics, il y a une option administrative qui n'est pas activée par défaut, mais qui peut l'être pour bloquer tous les accès par des robots et des araignées connus. L'option de filtrage des robots dans Google Analytics utilise une liste payante de filtres de robots fournis par l'Interactive Advertising Bureau (IAB), qui sont mis à jour systématiquement[40].

Il y a certains « bons robots » qui s'identifient clairement, comme les robots Google qui fouillent les sites Web pour créer des index de recherche. Cependant, le filtrage proactif des robots est problématique parce que certains sont programmés pour se comporter trompeusement comme des utilisateurs réguliers. Une approche fait appel au filtrage adaptatif, qui vise à améliorer progressivement la détection des robots et le filtrage par algorithme. Cependant, « plus le système de filtrage est raffiné, plus le risque est grand que les utilisateurs réguliers soient exclus avec les indésirables »[41], ce

---

[37] AlNoamany, Y. A., Weigle, M. C., & Nelson, M. L. (2013). Access patterns for robots and humans in web archives. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 339–348. https://doi.org/10.1145/2467696.2467722.

[38] Amshey, S. (n.d.) Bot shields: Activate! ensuring reliable repository download statistics. BePress. Webinar. Repéré à : https://www.bepress.com/webinar/bot-shields-activate-ensuring-reliable-repository-download-statistics/.

[39] IRUS-UK (2013). Position statement on the treatment of robots and unusual usage. Repéré à : https://irus.jisc.ac.uk/documents/IRUS-UK_position_statement_robots_and_unusual_usage_v1_0_Nov_2013.pdf.

[40] Moore, A. (2015). Eliminating bot traffic from Google Analytics once and for all. Repéré à : https://www.bounteous.com/insights/2015/04/01/eliminating-bot-traffic-google-analytics-once-and-all/?lang=en-ca.

[41] Information Power LTD. (2013). IRUS download data – identifying unusual usage. Repéré à : https://irus.jisc.ac.uk/documents/IRUS_download_data_Final_report.pdf.

qui n'est pas l'idéal pour les ressources ouvertes. Greene (2016) a procédé à un examen détaillé et complet de 10 approches différentes de la détection, pour conclure que « il n'y a pas de méthode capable de détecter avec exactitude tous les robots qui visitent un serveur Web donné », et fait plutôt valoir que l'objectif de ces techniques est de « capturer le plus grand nombre possible de robots tout en étiquetant comme robots le moins grand nombre de sessions humaines »[42].

L'utilisation de filtres pour les plateformes ouvertes de communication savante en est à ses premiers stades de l'expérimentation et de la compréhension. Avant la dernière publication de son code de pratique, COUNTER a formé un Groupe de travail sur les ROBOTS, qu'il a chargé d'étudier l'utilisation possible des filtres. Le Groupe de travail a essayé plusieurs techniques et décrit son travail comme « le début d'un ensemble de normes à développer pour aider à produire des statistiques d'utilisation convergentes, crédibles et comparables, qui pourront être agrégées sur de nombreux types de plateformes de communication savante »[43].

Une norme Web appelée protocole d'exclusion des robots est conçue pour informer les robots sur la façon d'interagir avec un site Web en signalant les pages à exclure de la recherche. Les robots ne sont pas tous conformes à la norme Web, et certains n'en tiendront pas compte, de sorte que cette méthode ne suffit pas en soi. La liste d'exclusion fournie par le fichier robots.txt peut être complétée par les plans de site, qui donnent une liste détaillée de la structure du site, ainsi que les zones signalées qui peuvent être incluses dans la recherche Web. Les plans de site aident à découvrir et à indexer le contenu d'un site, et l'existence d'une page Web bien structurée et facile à parcourir est un moyen connu d'accroître la probabilité d'un classement élevé de la page des résultats du moteur de recherche (SERP)[44]. Cela fait penser à une autre raison pour laquelle le blocage proactif des robots pourrait être désavantageux pour les objectifs des gestionnaires de dépôt de faire en sorte qu'il soit possible de découvrir le contenu de leurs DI.

Se posent également des questions importantes quant à savoir si le blocage des régimes d'utilisation automatisée peut exclure l'utilisation légitime; « les téléchargements automatisés ne sont pas nécessairement le fait de robots, p. ex., un établissement qui effectue une grande recherche documentaire utilise un script pour effectuer une recherche croisée dans un certain nombre de bases de données et de

---

[42] Greene, J. W. (2016). Web robot detection in scholarly open access institutional repositories. *Library Hi Tech, 34*(3), 500–520. https://doi.org/10.1108/LHT-04-2016-0048.

[43] Greene, J. W. (2017). Developing COUNTER standards to measure the use of open access resources. *Qualitative and Quantitative Methods in Libraries, 6*(2), 315-320. Repéré à : http://www.qqml-journal.net/index.php/qqml/article/view/410.

[44] Arlitsch, K., OBrien, P., & Rossmann, B. (2013) Managing search engine optimization: An introduction for library administrators (preprint), *Journal of Library Administration, 53*(2-3), 177-188. Repéré à : https://scholarworks.montana.edu/xmlui/handle/1/8671.

dépôts »[45]. Les tentatives visant à régler ce problème sont notamment l'ajout par COUNTER de la méthode d'accès « ECD », pour permettre aux créateurs de rapports de signaler les cas d'extraction de contenu et de données[46].

Indépendamment de ces questions compliquées, les robots sont à l'origine de près de 50 % du trafic Internet et de 85 % des téléchargements de DI[47] et il est au mieux de nos intérêts de continuer d'explorer des moyens de mieux différencier et de mieux définir les véritables régimes d'utilisation, afin de refléter plus fidèlement la valeur de nos dépôts.

Conclusion

Ce rapport a décrit le rôle important que les statistiques sur les dépôts institutionnels peuvent jouer dans l'évaluation de la recherche, mais surtout dans l'avancement et la compréhension de nos investissements continus envers le libre accès.

Il semble y avoir un niveau modeste à élevé d'effort et d'intérêt à l'échelle des établissements canadiens sondés pour suivre l'utilisation des données, et il est clair que les produits Google, tout comme les outils statistiques spécialisés fournis par les plateformes de dépôts institutionnels, sont les outils les plus privilégiés.

Notre exploration des nouveaux outils et des nouvelles approches pour l'amélioration et la normalisation des statistiques d'utilisation des dépôts institutionnels a révélé qu'il s'agit d'un domaine riche et en plein essor d'étude universitaire et technique qui va bien au-delà de ce que peut offrir une approche originale des statistiques. Le besoin de normes ouvertes et interopérables est manifeste, tout comme l'appel à protéger la vie privée des utilisateurs lorsqu'ils rassemblent des statistiques.

En ce qui concerne la convergence et la fiabilité, COUNTER est la seule norme internationalement reconnue pour la collecte et la publication de statistiques d'utilisation de la documentation savante diffusée sur le Web. COUNTER a également confié à des groupes de travail le soin de s'attaquer aux problèmes complexes associés à l'utilisation de robots, a créé et maintient le protocole SUSHI, qui permet la production de rapports par API, et a publié un code de pratique pour l'application des mêmes principes aux données de recherche.

---

[45] IRUS-UK (2013). Position statement… Ibid.

[46] Mellins-Cohen, (n.d.). The friendly guide to release 5: Technical notes for providers. Repéré à :. https://www.projectcounter.org/wp-content/uploads/2017/07/Tech_Notes_20170710.pdf.

[47] OBrien, P., Arlitsch, K. Mixter, J., Wheeler, J. & Sterman, L. (2017) "RAMP – the Repository Analytics and Metrics Portal"… Ibid.

Des nouveaux services de centralisation pour les statistiques des dépôts institutionnels offrent des outils et de l'expertise qui permettent d'accomplir les tâches de normalisation nécessaires en utilisant le code de pratique COUNTER. Grâce à cela, les dépôts institutionnels canadiens peuvent augmenter considérablement la fiabilité et la convergence de leurs statistiques.

Notre groupe a constaté que les statistiques IRUS-UK et OpenAIRE offrent les approches les plus favorables pour la collecte et la présentation des statistiques d'utilisation des dépôts institutionnels en raison de l'accent mis sur la production de statistiques conformes à COUNTER. Le projet RAMP n'a pas adopté une approche reconnue et normalisée de la collecte et de la publication des statistiques sur les dépôts institutionnels. Nous le félicitons pour les efforts qu'il a déployés pour créer un service de tableau de bord convivial qui lui est propre. Son intérêt pour l'établissement de renvois externes aux visionnements de PDF est nouveau et mérite d'être étudié. Bien qu'ils déploient des approches différentes, IRUS-UK et OpenAIRE sont tout aussi capables de répondre aux exigences que nous avons établies. IRUS-UK représente une solution de rechange valable qui mériterait tout autant d'être considérée qu'un service normalisé à tous les dépôts institutionnels canadiens.

Enfin, nous recommandons que les dépôts institutionnels canadiens choisissent d'adopter collectivement les statistiques OpenAIRE, en raison d'une part du fait de l'utilisation de l'outil d'analyse Web Matomo axé sur la protection des renseignements personnels, et d'autre part de son approche globale du travail avec les dépôts ouverts. Le choix s'harmonise également avec d'autres activités du Groupe de travail sur les dépôts ouverts de l'ABRC, qui poursuit ses activités d'intégration aux initiatives internationales de diffusion du libre accès.

Annexe 1 : Questions de l'enquête

Nom du dépôt || Name of your Institution

Quel est l'URL de votre dépôt? || What is your Repository's URL?

Suivez-vous l'utilisation de votre dépôt? || Are you Tracking Repository Usage?

Est-ce que l'analyse (« analytics ») est comprise dans le dépôt, ou utilisez-vous un service externe? || Are analytics provided internally by IR, or by external service?

Vous servez-vous de Google Analytics pour votre dépôt? || Do you enable Google Analytics for your IR?

Est-ce que votre bibliothèque utilise Google Search Console? || Does your library use Google Search Console?

Contribuez-vous les plans de site de votre dépôt à Google? || Do you submit sitemaps of your IR to Google?

Est-ce que votre dépôt emploie des moyens pour bloquer les robots (« bots »)? || Does your IR employ any attempts to block bot traffic?

Qu'est-ce que vous désirez savoir au sujet de l'utilisation de votre dépôt? || What would you like to know about your IR's usage?

Est-ce que vous recueillez et utilisez les fichiers journaux (« log files »)? || Do you collect and utilize log files?

Diffusez-vous les statistiques de votre dépôt avec les membres de votre campus? || Do you share repository statistics with your campus?

D'après vous, quel est le niveau de satisfaction de vos usagers quant aux statistiques que vous diffusez? || What is your perception of your user's satisfaction with the statistics you provide?