Indicateurs de consultation, indicateurs stratégiques : leur production à partir de l'analyse des consultations des documents d'une archive en libre accès

Accesses indicators, strategic indicators: their production from log files analysis of an open access archive

ROSSI Pier Luigi
Institut de recherche pour le développement (IRD)
32, avenue Henri Varagnat
93140 Bondy - France
Mail: rossi@ird.fr

Résumé

L'archive documentaire de l'IRD (Institut de recherche pour le développement) est constituée par l'ensemble des productions scientifiques de l'Institut. Actuellement plus de 66.500 documents sont disponibles gratuitement sur Internet. En associant, en mode relationnel, la base de données bibliographiques et celle concernant les consultations de ces documents (« fichiers log Apache »), nous pouvons produire des tableaux de bord dynamiques en fonction de plusieurs critères et de plusieurs indicateurs. Nous allons présenter plusieurs indicateurs originaux ainsi que nos méthodes d'analyse et de production.

Mots-clés. Indicateurs, accès internet, libre accès, fichiers journaux, analyses statistiques.

Abstract

The IRD (French National Research Institute for Sustainable Development) archive contains all the scientific outputs of the Institute. Currently more than 66,500 documents are freely available on the Internet. By associating, in relational mode, the bibliographic database and accesses database ("Apache log files"), we can produce dynamic dashboards according to several criteria and several indicators. We will present several original indicators, our analysis methods and indicators production.

Keywords. Indicators, internet access, free access, log files, statistical analysis.

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

Introduction

L'IRD (Institut de recherche pour le développement)[1] est un institut de recherche français au service des pays méditerranéens et intertropicaux qui font de la science et de l'innovation l'un des premiers leviers de leur développement. Il s'agit d'un établissement public français à caractère scientifique et technologique (EPST) placé sous la double tutelle du ministère de l'Enseignement supérieur et de la Recherche et du ministère des Affaires étrangères et du Développement international[2].

Le décret du 1er décembre 1955 (ORSTOM, 1955) organisant les services de l'Office de la recherche scientifique et technique outre-mer, publié au Journal officiel de la République française du 21 décembre 1955, définit l'existence d'un centre de documentation rattaché à la Direction de l'Institut. L'article 6 stipule que « le Centre de documentation est chargé d'assurer le dépouillement, la conservation et la diffusion de la documentation scientifique et technique se rapportant aux activités de l'O.R.S.T.O.M. ».

Depuis 1955, l'Institut a donc créé une archive institutionnelle avec la volonté de préserver et de diffuser ses productions scientifiques. Ce fonds documentaire patrimonial se compose actuellement (aout 2018) de 98.400 documents et constitue le fonds documentaire de l'IRD (FDI).

Cette collection a été informatisée depuis 1986 avec la création de la base de données bibliographiques Horizon (Roux-Fouillet, 1988). Dès 1996, nous avons initié sa numérisation (Rossi, 1997) et 66.500 documents sont actuellement (aout 2018) disponibles gratuitement sur Internet (Rossi et Ngoma-Mouaya, 2000).

Compte tenu des implantations géographiques des centres de recherche et des missions des chercheurs, tout au long de l'histoire de l'Institut, une partie significative des travaux scientifiques réalisés (plus de 39.000 documents en accès libre) concerne les pays du continent africain.

Fichiers journaux[3] et filtrage de données

Le site hébergeant l'archive institutionnelle de l'IRD fonctionne avec Apache HTTP Server[4]. Ce logiciel stocke de manière détaillée

[1] Voir : http://www.ird.fr/l-ird/presentation (Page consultée le 6 août 2018).

[2] L'IRD a été créé en 1943 sous le nom d'Office de la recherche scientifique coloniale (ORSC). En 1953, il est devenu Orstom (Office de la recherche scientifique et technique d'outre-mer). En 1984, il prend le nom d'Institut français de recherche scientifique et technique pour le développement en coopération, tout en conservant son acronyme et devient IRD en 1998. Voir : http://www.ird.fr/l-ird/historique (Page consultée le 6 août 2018).

[3] Voir : https://fr.wikipedia.org/wiki/Historique_(informatique) (Page consultée le 7 août 2018).

[4] Voir : https://httpd.apache.org/ (Page consultée le 7 août 2018).

l'historique des consultations concernant les échanges de fichiers qui se réalisent entre le serveur et le client (la machine qui a consulté les fichiers disponibles sur le serveur) (Agosti *et al.*, 2012).

154.126.12.202 - - [**12/Feb/2017:17:50:33 +0100**] "GET **/exl-doc/pleins_textes/pleins_textes_5/b_fdi_08-09/10230.pdf** HTTP/1.1" **200** 2300612 "http://**www.google.fr**/search?ei=bZGgWJHPIIzhvATnxpywDQ&q=**trachysphaera+fructigena+pdf**&btnG=" "Mozilla/5.0 (Linux; U; Android 4.2.2; en-US; HTC Butterfly Build/JDQ39) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 UCBrowser/11.1.5.890 U3/0.8.0 **Mobile Safari**/534.30"

Figure 1. *Format du fichier d'historique, en mode « combiné »*

Le format du fichier d'historique, en mode « combiné »[5], contient plusieurs informations (figure 1) incluant: l'adresse IP du client (154.126.12.202), l'heure à laquelle la demande a été reçue (12/Feb/2017:17:50:33 +0100), l'URL du fichier consulté (/exl-doc/.../10230.pdf), le code d'état renvoyé par le serveur (200), le site depuis lequel le client a lancé sa requête[6] (http://www.google.fr ...), la question éventuelle posée par le client (trachysphaera fructigena pdf), le type d'appareil utilisé (... Mobile ...) et le navigateur utilisé par le client (... Safari ...).

Pour mener à bien notre étude, nous avons géolocalisé les adresses IP en utilisant la base de données GeoLite Country de MaxMind[7]. Dans l'exemple de la figure 1, l'adresse IP est située à Madagascar.

Le fichier journal est filtré (Krishnagandhi et Dhas, 2016) en conservant uniquement les lignes ayant accès aux fichiers PDF avec le statut « 200 »[8]. Le fichier résultant est ensuite décomposé en plusieurs sous-fichiers. Le champ contenant le type de navigateur inclut également la signature des robots de recherche. Cela permet de générer des fichiers journaux relatifs à l'indexation effectuée par les moteurs de recherche, avec notamment le moteur de recherche Google (figure 2).

66.249.64.221 - - [07/Feb/2017:12:01:04 +0100] "GET /exl-doc/pleins_textes/divers17-01/010068365.pdf HTTP/1.1" 200 631342 "-" "Mozilla/5.0 (compatible; **Googlebot**/2.1; +http://www.google.com/bot.html)"

Figure 2. *Indexation d'un fichier effectuée par « Googlebot »*

Nous rejetons et générons un fichier "spam" avec toutes les lignes du fichier journal qui masquent par un tiret ("-") le champ référent (le champ du site à partir duquel le client a lancé sa requête) ou le champ navigateur : nous présupposons que les requêtes qui masquent ces deux champs peuvent être assimilées soit à une attaque par déni de

[5] Voir : https://httpd.apache.org/docs/2.4/fr/logs.html (Page consultée le 7 août 2018).
[6] Voir : https://fr.wikipedia.org/wiki/Référent_(informatique) (Page consultée le 7 août 2018).
[7] Voir : https://dev.maxmind.com/geoip/legacy/geolite/ (Page consultée le 7 août 2018).
[8] Voir : https://tools.ietf.org/html/rfc2616#page-58 (Page consultée le 7 août 2018). Le statut « 200 » du code http indique que la requête a été traitée avec succès.

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

service[9] soit à un « referer spam »[10]. Selon la même logique, on rejette également les accès provenant de la même adresse IP et consultant le même fichier qui sont réalisés dans un intervalle inférieur à 5 minutes.

Le fichier de travail filtré (Jansen, 2006) qui se compose de tous les accès « validés » (« human access ») contient 2.748.000 lignes, soit 39% du fichier brut d'origine. Le fichier des « accès robots » et le fichier des « accès spam » représentent respectivement 42% et 19% du fichier brut d'origine.

Répartition géographique des accès

Après ces différentes étapes de filtrage (Kaur et Aggarwal, 2017), nous avons un fichier constitué (pour l'essentiel) de « human access ». Il est prêt à produire des indicateurs de statistiques et d'accès (Martinez-Comeche, 2017 ;Rossi et Thiaw, 2012 ; Rossi *et al.*, 2018).

Comme tous les accès ont été géolocalisés, il est possible de produire des statistiques en fonction de la répartition géographique par pays et par continents. Nous présentons ces résultats à l'aide de graphiques «treemap»[11] qui nous ont parus bien adaptés à la présentation de ce type de données (figure 3).

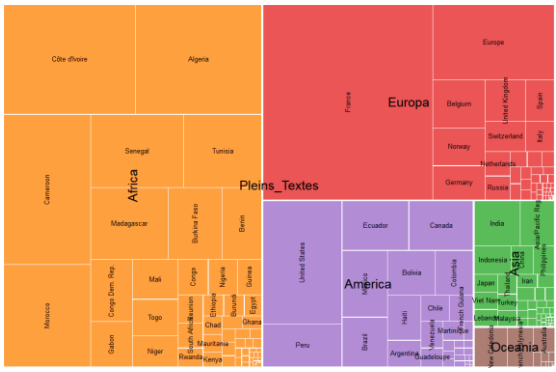

Figure 3. *Géolocalisation des accès*

Si l'on considère les données « validées » de l'année 2017[12], les accès sont situés dans 234 pays dont 56 pays africains. La France est le premier pays où les consultations sont localisées avec 459.021 accès, soit 17% du total. Pour les pays du continent africain, on compte 1.290.723 accès, 47% du total (Europe: 29%, Amériques: 18%, Asie: 5.1%, Océanie: 1.6%). La Côte d'Ivoire est le premier pays d'Afrique avec 199.424

[9] Voir https://fr.wikipedia.org/wiki/Attaque_par_déni_de_service (Page consultée le 7 août 2018).

[10] Voir https://fr.wikipedia.org/wiki/Referer_spam (Page consultée le 7 août 2018).

[11] Voir : https://fr.wikipedia.org/wiki/Treemap (Page consultée le 7 août 2018). Voir également : Treemaps for Space-Constrained Visualization of Hierarchies: The History of Treemap Research at the University of Maryland. http://www.cs.umd.edu/hcil/treemap-history/index.shtml (Page consultée le 7 août 2018).

[12] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/2017.htm (Page consultée le 9 août 2018). Le tableau de bord contient, entre autre, des informations sur les répartitions et les fréquences des langues des documents consultés. On consultant le tableau on constate que **les documents en espagnol sont deux fois plus consultés** que les documents en français et que **ceux en anglais sont deux fois moins consultés** que ceux en français.

accès, soit 7,3% du total. Les pays du Maghreb (Algérie, Maroc et Tunisie) génèrent 392.516 accès, soit 14% du total.

Le Nigéria est le premier pays anglophone d'Afrique avec 14.201 accès, soit 0,5% du total. Les États-Unis sont le premier pays des Amériques avec 134.998 accès, soit 4,9% du total. L'Inde est le premier pays d'Asie avec 28.800 accès, soit 1,1% du total. La Nouvelle-Calédonie est le premier pays d'Océanie avec 18.406 accès, soit 0,7% du total.

Ces statistiques peuvent témoigner des contextes géopolitiques tendus caractérisant certains pays africains. Notre Institut a, par exemple, produit un grand nombre de publications sur le Tchad (1.344 documents) et la République centrafricaine (787 documents) qui sont librement accessibles. Le nombre de consultations dans ces pays est relativement faible: pour le Tchad, on compte 5.110 accès en 2017, soit 0,19% du total et pour la République centrafricaine 3.560 accès, soit 0,13% du total. Ces résultats s'expliquent par la prise en compte du nombre d'internautes par rapport à la population du pays (pénétration internet), qui est respectivement de 5% pour le Tchad et de 5,4% pour la République centrafricaine (Internet World Stats 2018). Ces éléments factuels sont amplifiés par la forte instabilité politique et sociale qui caractérise ces pays depuis plusieurs années. Le prix de l'accès à Internet est également un facteur limitant dans plusieurs pays africains (Cable.co.uk 2017).

L'analyse de géolocalisation des accès peut être croisée avec la langue des documents. Le graphique des consultations des documents en français[13] est visuellement semblable au graphique de la figure 3, mais les pays francophones sont caractérisés par des pourcentages plus élevés. On compte 2.217.559 accès à des documents en français (ce qui représente 80,7% des de l'ensemble des consultations), situés dans 224 pays dont 55 pays africains. La France est le premier pays où les consultations sont localisées avec 447.349 accès, soit 20.2% du total. Pour les pays du continent africain, on compte 1.243.482 accès, 56% du total (Europe: 32%, Amériques: 8,2%, Asie: 1,9%, Océanie: 1,6%). La Côte d'Ivoire est le premier pays d'Afrique avec 198.283 accès, soit 8,9% du total. Les pays du Maghreb (Algérie, Maroc et Tunisie) génèrent 388.633 accès, soit 18% du total. Ce filtrage de données (langue du document : français), fait apparaître Haïti (0,9%, 20866 accès) dans les Amériques et le Liban (0,2%, 5302 accès) en Asie.

En ce qui concerne les documents en anglais (figure 4)[14], on compte 326.351 accès, situés dans 228 pays, dont 55 pays africains. L'Asie est le premier continent avec 94.952 accès, soit 29% du total. L'Inde représente 8,2% du total avec 26.667 accès.

Les États-Unis sont le premier pays où les accès sont situés avec 37.244 accès, soit 11,4% du total.

[13] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/FR-2017.html (Page consultée le 9 août 2018).

[14] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/EN-2017.html (Page consultée le 9 août 2018).

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

Pour les pays du continent africain, on dénombre 71.435 accès, 22% du total (Europe: 24%, Amériques: 22%, Océanie: 3,3%). Le Nigeria est le premier pays d'Afrique avec 10.104 accès, soit 3,1% du total.

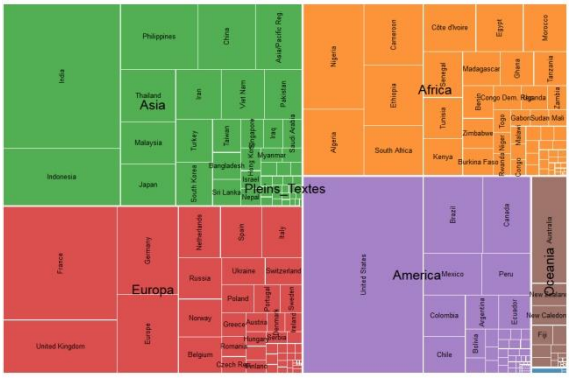

Figure 4 : *Accès aux documents en anglais*

En ce qui concerne les documents en espagnol (figure 5)[15], on compte 234.059 accès, situés dans 161 pays. Les Amériques sont le premier continent avec 221.135 accès, soit 94% du total. L'Equateur est le premier pays avec 43.295 accès, soit 18% du total. Les pays andins les moins avancés (Equateur, Pérou et Bolivie) génèrent 113.344 accès, soit 48% du total. L'Espagne est le premier pays d'Europe avec 6.368 accès, soit 2,7% du total.

Figure 5 : *Accès aux documents en espagnol*

Ces résultats sont révélateurs de l'impact des productions de l'IRD, publiées en espagnol, dans les pays d'Amérique latine : en 2017 2.241 fichiers en libre accès ont été consultés au moins une fois et la moyenne des consultations est de 104, une moyenne 4 fois plus importante que celle des consultations des documents en langue anglaise.

Tableaux de bord[16], construction et indicateurs

En croisant de façon relationnelle la base de données bibliographiques et celles des consultations, il est possible de générer de façon

[15] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/ES-2017.htm (Page consultée le 9 août 2018).

[16] Les tableaux de bord sont réalisés en mode dynamique avec des formulaires au format html intégrant des scripts en php et des requêtes vers la base de données sql contenant les tables des statistiques de consultations et des métadonnées bibliographiques des documents du fonds documentaire de l'IRD. Les graphes sont réalisés avec des bibliothèques javascript. Pour les graphes treemap nous utilisons « protovis » (voir : https://github.com/mbostock/protovis, page consultée le 9 août 2018) et pour les autres graphes nous utilisons « highcharts » (voir : https://www.highcharts.com/, page consultée le 9 août 2018).

dynamique des tableaux de bord en fonction de plusieurs critères : auteur, année de publications, langue du document, thématique des documents, pays à l'origine des consultations.

Ces tableaux de bord sont constitués de plusieurs éléments. Si l'on considère le tableau de bord du « thème auteur »[17] (analyse des consultations des documents en libre accès d'un auteur), le premier élément est un tableau qui résume les principales données.

Indicateur	Valeur
Nombre de fichiers « Pleins_Textes » consultés	350
Nombre de consultations « Pleins_Textes »	27989
Nombre moyen de consultations par fichier « Pleins_Textes »	79,97
Facteur D (nombre de documents consultes au moins "n" fois)	75
Somme des consultations des Top 75	24120
Consultations Top 75 sur le total des consultations des documents de Roose, Eric	86,2%

Figure 6 : *Résumé des principaux indicateurs d'un tableau de bord*

Pour cet exemple (figure 6), au cours de l'année 2017, 350 différents fichiers du même auteur (Roose, Eric) ont été consultés au moins une fois. Pour ces documents, 75 ont été consultés au moins 75 fois au cours de l'année 2017. Cette valeur, indiquant le nombre de documents consultés au moins n fois (calculée avec un script en php), détermine le « Facteur D »[18] de l'auteur pour la période annuelle des consultations.

Dans le tableau de bord de l'auteur on affiche ensuite le tableau des

Tableau des Top 75 (nombre de fichiers = Facteur D) (2017)

Classement	Nombre de consultations	Premier pays consultant	Indice d'appropriation par le premier pays consultant	Référence bibliographique
Top 1	2319	Cameroon (291)	0.13	Roose, Eric. Causes et facteurs de l'érosion hydrique sous climat tropical : conséquences sur les méthodes antiérosives. 1984 (Sc. Terre)
Top 2	1617	Côte d'Ivoire (1128)	0.70	Roose, Eric; Chéroux, Michel; Humbel, François-Xavier (collab.); Perraud, Alain (collab.). Les sols du bassin sédimentaire de Côte d'Ivoire. 1966 (Sc. Terre)
Top 3	1323	Cameroon (174)	0.13	Roose, Eric (ed.); Duchaufour, H. (ed.); De Noni, Georges (ed.). Lutte antiérosive : réhabilitation des sols tropicaux et protection contre les pluies exceptionnelles. 2012 (Sc. Terre)
Top 4	1051	Côte d'Ivoire (179)	0.17	Roose, Eric. Impact du défrichement sur la dégradation des sols tropicaux. 1984 (Sc. Terre)
Top 5	955	Morocco (571)	0.60	Roose, Eric (ed.); Sabir, M. (ed.); Laouina, A. (ed.); Benchakroun, F. (collab.); Al Karkouri, J. (collab.); Lauri, P. (collab.); Qarro, M. (collab.). Gestion durable des eaux et des sols au Maroc : valorisation des techniques traditionnelles méditerranéennes. 2010 (Sc. Terre)
Top 6	762	Benin (141)	0.19	Roose, Eric. Dégradation des terres et développement en Afrique de l'Ouest. 1985 (Sc. Terre)
Top 7	734	France (401)	0.55	Roose, Eric; Blancaneaux, Philippe; Freitas, P.L. de. Un simple test de terrain pour évaluer la capacité d'infiltration et le comportement hydrodynamique des horizons pédologiques superficiels : méthode et exemples. 1993 (Sc. Terre)

Figure 7 : *Extrait du tableau des documents les plus consultés*

« n » documents les plus consultées (la valeur de n est égale au Facteur D). Pour cet exemple (figure 7), le document « Top 1 » à été consulté 2.319 fois en 2017, le pays ayant le plus consulté ce document (premier

[17] Le tableau de bord complet est accessible à http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/Roose-2017.htm (Page consultée le 7 août 2018).

[18] Cet indicateur, Facteur D, ou Downloading factor, s'inspire directement de l'indice H. Voir : https://fr.wikipedia.org/wiki/Indice_h (Page consultée le 7 août 2018).

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

pays consultant) est le Cameroun (291 accès) et l'indice d'appropriation du Cameroun est de 0,13 (soit 13% de l'ensemble des consultations). Si l'on parcourt ce tableau, on remarque que le deuxième document porte sur la Cote d'Ivoire (voir le titre de la référence bibliographique) : dans ce cas l'indice d'appropriation du premier pays consultant (Cote d'Ivoire) est 0,70.

L'observation complète du tableau des documents les plus consultés, montre que si un document porte sur un pays spécifique (dont le nom apparait généralement dans le titre du document), « le premier pays consultant » est le pays indiqué dans le titre et son indice d'appropriation est élevé.

Par ailleurs, ce tableau permet d'observer que le nombre des consultations est indépendant de l'année de publication des documents : le déterminant des consultations, et, par conséquence, de leur nombre, est le contenu du document ainsi que son titre. Les pages de résultats des moteurs de recherche[19] qui s'affichent après une requête comportent le titre des documents trouvés et ceux-ci déterminent, très probablement, les choix de consultation des internautes.

Le tableau de bord « auteur », comporte ensuite le graphe « treemap » avec la répartition géographique totale des consultations. Dans le cas spécifique, avec les outils intégrés d'analyse dynamique du graphe on constate que les consultations venant de l'Afrique constituent 17.509 accès (63% du total), reparties sur 52 pays.

Le nombre total des fichiers consultés et le nombre total des consultations sont ensuite repartis en fonction de la langue des documents et représentés sur des graphes. La comparaison de ces valeurs permet de générer un indice de spécialisation en fonction de la

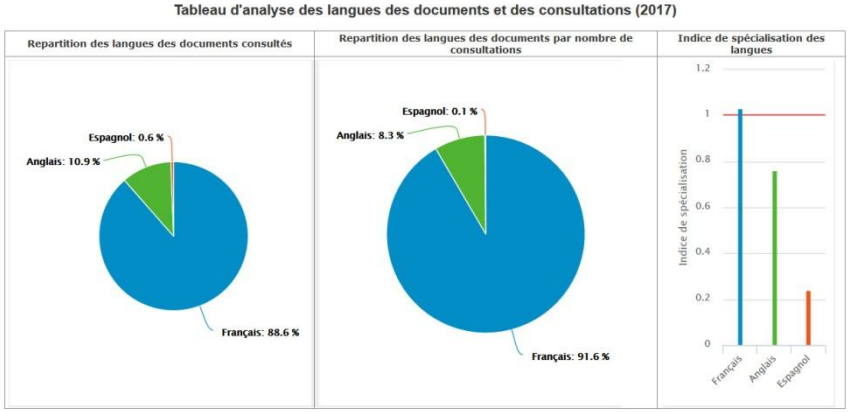

Figure 8 : *Répartition des accès en fonction de la langue des documents*

langue. Pour l'exemple choisi (figure 8), on constate ainsi que les

[19] Il est à noter que Google est à l'origine de plus de 85% des requêtes faites par les internautes vis-à-vis de notre serveur. Ce résultat est obtenu par le comptage de champ « referer » dans la table des accès « validés » de l'année 2017.

documents en français sont légèrement sur-consultés (1,03) et que les documents en anglais (0,76) et en espagnol (0,24) sont sous-consultés.

Le tableau de bord intègre un graphe dynamique qui permet de calculer et visualiser l'indice de Pareto[20]. Pour cet exemple on constate qu'il suffit de considérer 16% des fichiers le plus consultés pour atteindre 80% des consultations.

Répartitions thématiques des consultations des pays

Dans la base de données bibliographique du fonds documentaire de l'IRD, chaque document est associé à un thème (indexation des traitements effectuée par les documentalistes de l'IRD). Les thèmes de la base de données bibliographiques sont les suivants : Sciences de l'ingénieur (Sc. Ing.), Océanographie-Hydrobiologie (Oceano + Hydrobio.), Santé (Santé), Sciences de la Terre (Sc. Terre), Sciences végétales et animales (Sc. Veg + Ani), Sciences humaines et sociales (Sc. Hum. + Soc.).

Les données d'accès peuvent être filtrées par rapport à un pays et associées aux thèmes des documents. Ces données "pays" peuvent être comparées à la distribution globale des accès par thème. Ceci montre un index d'intérêt (indice de spécialisation thématique) pour les thèmes par rapport à chaque pays (figure 9).

Pour l'Afrique du Sud[21], les documents concernant l'Océanographie-Hydrobiologie et les Sciences végétales et animales sont sur-consultés. Les documents concernant les Sciences de l'ingénieur et les Sciences de la Terre sont sous-consultés.

Pour l'Algérie[22], les documents concernant les Sciences de l'ingénieur, les Sciences de la Terre et les Sciences végétales et animales sont sur-consultés. Les documents concernant les Sciences humaines et sociales sont sous-consultés.

Pour la Côte d'Ivoire[23], les documents concernant les Sciences humaines et sociales sont sur consultés. Les documents concernant tous les autres sujets sont sous-consultés.

Pour l'Egypte[24], les documents concernant l'Océanographie-Hydrobiologie et les Sciences végétales et animales sont sur-consultés. Les documents concernant les Sciences de l'ingénieur et les Sciences humaines et sociales sont sous-consultés.

---

[20] Voir : https://fr.wikipedia.org/wiki/Indice_de_Pareto (Page consultée le 10 août 2018).

[21] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/ZAF-2017.htm (Page consultée le 9 août 2018).

[22] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/DZA-2017.htm (Page consultée le 9 août 2018).

[23] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/CIV-2017.htm (Page consultée le 9 août 2018).

[24] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/EGY-2017.htm (Page consultée le 9 août 2018).

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

Pour le Maroc[25], les documents concernant les Sciences de l'ingénieur et les Sciences de la Terre sont sur-consultés. Les documents concernant

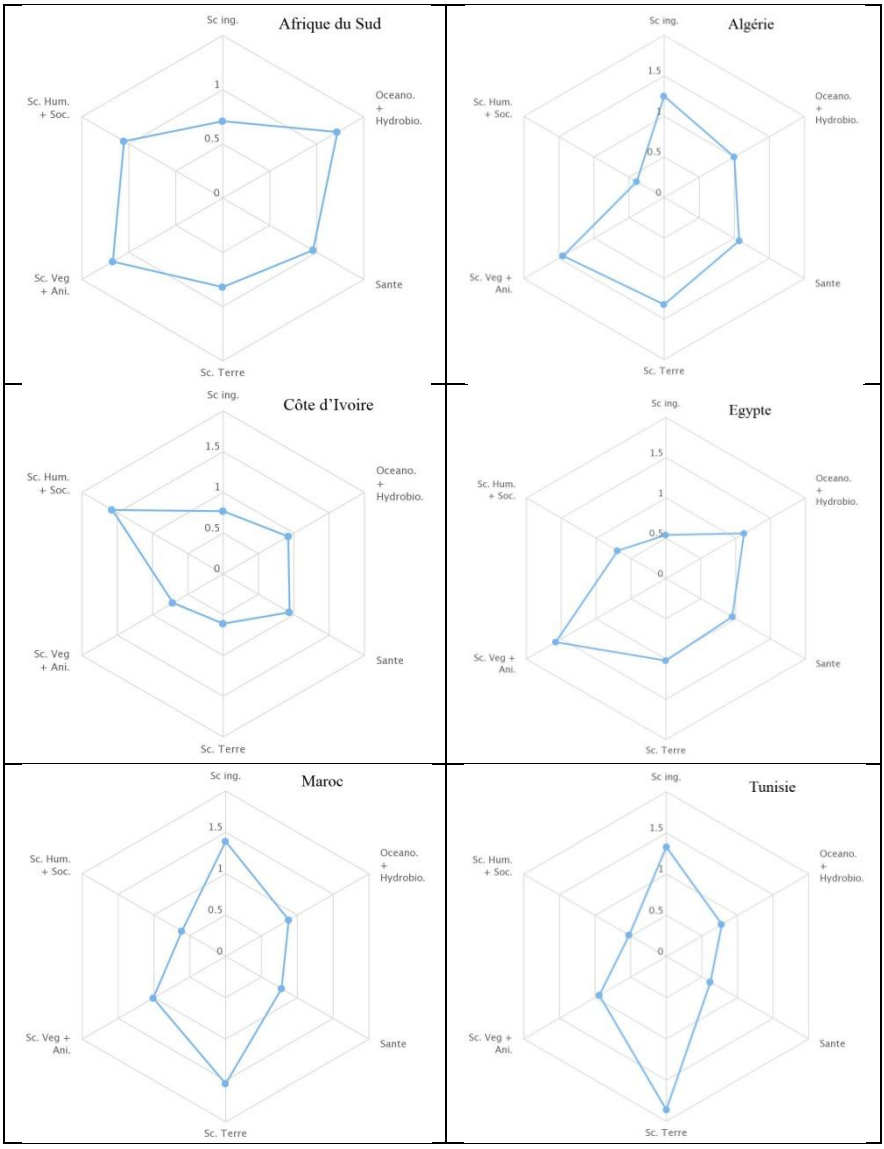

Figure 9. *Répartitions thématiques des consultations de 6 pays d'Afrique*

les Sciences humaines et sociales, l'Océanographie-Hydrobiologie et la Santé sont sous-consultés.

[25] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/MOR-2017.htm (Page consultée le 9 août 2018).

Pour la Tunisie[26] la répartition des thématiques des consultations est comparable à celle du Maroc. On remarque néanmoins un très fort intérêt pour les documents concernant les Sciences de la Terre.

En analysant les profils disciplinaires de ces pays, on remarque que les graphes de l'Afrique du Sud et de l'Egypte (pays essentiellement anglophones) sont semblables. Il en est de même pour le Maroc et la Tunisie avec des intérêts très forts pour les Sciences de l'ingénieur et les Sciences de la Terre.

Pédologie et Hydrologie

Au cours de 2016 et 2017, nous avons mené une campagne de numérisation pour traiter toutes nos productions scientifiques concernant la Pédologie (Fargier 2015) et l'Hydrologie. Ce sont deux disciplines qui ont caractérisé l'histoire scientifique de notre institut et qui ont permis de produire un nombre important de publications particulièrement significatives pour les pays dans lesquels elles ont été réalisées : aussi bien en Afrique qu'en Amérique latine.

En 2017, 6.696 documents concernant la Pédologie ont été consultés au moins une fois et le nombre total des accès est de 331.600[27]. Les accès se repartissent dans 213 pays dont 56 pays africains. Les documents ont donc été consultés en moyenne 50 fois, pour toute l'année 2017. La France est le premier pays où les consultations sont localisées avec 55.813 accès, soit 17% du total. Pour les pays du continent africain, on compte 179.310 accès, 54% du total (Europe : 28%, Amériques : 13%, Asie : 4,2%, Océanie : 0,7%). L'Algérie est le premier pays d'Afrique avec 39.048 accès, soit 12% du total. Les pays du Maghreb (Algérie, Maroc et Tunisie) génèrent 83.392 accès, soit 25% du total. Six documents, dont 3 portant sur les techniques d'analyse des sols, totalisent en 2017 plus de 3.000 accès chacun.

En 2017, 6.242 documents concernant l'Hydrologie ont été consultés au moins une fois et le nombre total des accès est de 212.895[28]. Les accès se repartissent dans 200 pays, dont 55 pays africains. Les documents ont donc été consultés en moyenne 34 fois pour toute l'année 2017. La France est le premier pays où les consultations sont localisées avec 34.274 accès, soit 16% du total. Pour les pays du continent africain, on comptabilise 105.605 accès, 50% du total (Europe : 26%, Amériques : 20%, Asie : 3,2%, Océanie : 0,8%). L'Algérie est le premier pays d'Afrique avec 20.409 accès, soit 9,6% du total. Les pays du Maghreb (Algérie, Maroc et Tunisie) génèrent 49.976 accès, soit 23% du total. Le Brésil est le premier pays des Amériques avec 7.970 accès, soit 3,7% du total. Trois documents concernant les

[26] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/TUN-2017.htm (Page consultée le 9 août 2018).

[27] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/Pedologie-2017.htm (Page consultée le 9 août 2018).

[28] Le tableau de bord complet est accessible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/acces/Hydrologie-2017.htm (Page consultée le 9 août 2018).

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

techniques d'analyse hydrologique et la construction de barrages, totalisent en 2017 plus de 3.000 accès chacun.

L'indice de Pareto est de 11.9% pour la Pédologie et de 12.1% pour l'Hydrologie, ce qui signifie qu'il faut prendre en considération environ 12% des documents les plus consultés pour atteindre 80% des consultations. Ces deux valeurs, tout à fait comparables, sont largement en dessous de 20%. Ainsi les consultations des documents se concentrent sur une part « relativement » faible des documents disponibles.

L'indice de spécialisation des consultations par rapport aux langues que nous avons mis au point compare les pourcentages du nombre des documents consultés par rapport au nombre total des consultations des documents en fonction de la langue. Les tableaux de bord de ces deux disciplines fournissent des informations intéressantes par rapport à l'analyse des langues des documents en fonction des fréquences des consultations.

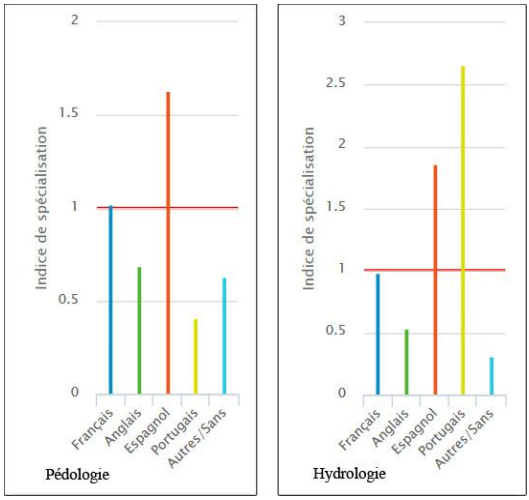

Figure 10 : *Indice de spécialisation de langues des documents de Pédologie et d'Hydrologie en fonction des accès*

La figure 10 illustre les résultats obtenus par rapport à ces deux disciplines. Pour la Pédologie on observe que les documents en anglais (IS = 0,69) et en portugais (IS = 0,41) sont sous-consultés et que les documents en espagnol sont sur-consultés (IS = 1,63). Pour l'Hydrologie on observe que les documents en anglais (IS = 0,54) sont sous-consultés et que les documents en espagnol (IS = 1,86) et en portugais (IS = 2.66) sont sur-consultés.

Conclusions

Le projet de numérisation que nous avons lancé en 1996 a permis la numérisation et la mise en libre accès d'environ 66% de la production scientifique de l'IRD (plus de 66.500 documents en format pdf). Le projet se poursuit et environ 4.000 fichiers pdf par an s'ajoutent à l'archive ouverte.

L'analyse des statistiques de consultation de nos documents en libre accès montre l'impact des documents produits par l'IRD vis-à-vis des utilisateurs de l'internet, notamment ceux des pays en développement et des pays émergents. L'association, en mode relationnel, des données de consultation et des métadonnées décrivant les documents

disponibles permettent de réaliser des tableaux de bord relativement élaborés et diversifiés : auteur, année de publication, langue du document, thématique des documents, pays à l'origine des consultations.

Dans ces tableaux de bord nous pouvons introduire des indicateurs originaux et spécifiques aux types de données dont nous disposons : « downloading factor », indice d'appropriation, indice de spécialisation en fonction de la langue, indice de spécialisation relatif à la thématique du document.

La combinaison de ces données avec les thèmes et les langues des documents montre des spécificités pour chaque pays (thématique) ou pour les aires linguistiques (pays francophones versus pays anglophones) du continent africain.

L'analyse montre que les documents les plus consultés sont souvent des documents publiés dans les années 90 sur les méthodologies, les techniques d'analyse et la construction de structures de génie civil (barrages).

La langue des documents apparait comme un facteur déterminant des consultations en fonction des pays : dans les pays francophone d'Afrique l'essentiel des consultations concernent des documents en français. Une vision stratégique des publications scientifiques vis-à-vis de ces pays devrait éventuellement prendre en compte ce facteur de la langue plutôt que de privilégier « systématiquement » des choix de publication en langue anglaise dans des revues à fort facteur d'impact.

Puisque l'on peut constater que les répartitions thématiques des consultations de chaque pays ne sont pas homogènes, le choix et les orientations des recherches pourraient en tenir compte pour privilégier des thématiques sur-consultées ou, à contrario, pour produire des publications scientifiques pouvant créer ou développer des intérêts plus originaux.

En tout état de cause, nous n'envisageons pas l'utilisation de ces résultats à des fins d'évaluation (des individus, des laboratoires, des institutions)[29] mais plutôt comme des éléments permettant d'élaborer des réflexions afin de comprendre l'utilisation de nos productions scientifiques par les internautes ainsi que l'impact qu'elles peuvent avoir, notamment dans les pays des Suds.

Bibliographie

AGOSTI, Maristella ; CRIVELLARI, Franco, DI NUNZIO, Giorgio Maria (2012). Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. In *Data*

[29] Une littérature très copieuse existe sur les questions de bibliométrie (voir : https://fr.wikipedia.org/wiki/Bibliométrie, page consultée le 9 août 2018), de scientométrie (voir : https://fr.wikipedia.org/wiki/Scientométrie, page consultée le 9 août 2018) et d'altmetrics (voir : https://en.wikipedia.org/wiki/Altmetrics, page consultée le 9 août 2018) et les usages qui sont faits en matière d'évaluation.

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

Mining and Knowledge Discovery, vol. 24, n° 3, pp. 663-696. https://doi.org/10.1007/s10618-011-0228-8

Cable.co.uk (2017). *Study of broadband pricing in 196 countries reveals vast global disparities in the cost of getting online*. [En ligne]. Disponible à : https://www.cable.co.uk/about/media-centre/releases/new-worldwide-broadband-price-league-unveiled/ (Page consultée le 10 août 2018).

FARGIER, Nathalie (2015) Numériser la littérature grise scientifique. In *I2D Information, données et documents*, vol. 52, n° 1, pp. 61-62. [En ligne]. Disponible à : https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-1-page-61.htm (Page consultée le 10 août 2018).

Internet World Stats (2018). *Internet Users Statistics for Africa*. [En ligne]. Disponible à : https://www.internetworldstats.com/stats1.htm (Page consultée le 10 août 2018).

ITU/UNESCO Broadband Commission for Sustainable Development (2017). *The state of broadband 2017: broadband catalyzing sustainable development*. ITU, Unesco, 2017, 104 p. [En ligne]. Disponible à : https://www.itu.int/dms_pub/itu-s/opb/pol/S-POL-BROADBAND.18-2017-PDF-E.pdf (Page consultée le 10 août 2018).

JANSEN, Bernard J. (2006). Search log analysis: what it is, what's been done, how to do it. In *Library & Information Science Research*, vol. 28, n° 3, pp. 407-432. http://dx.doi.org/10.1016/j.lisr.2006.06.005

KAUR, Navjot; AGGARWAL, Himanshu (2017) A Novel Semantically-Time-Referrer based Approach of Web Usage Mining for Improved Sessionization in Pre-Processing of Web Log. In *International journal of advanced computer science and applications*, vol. 8, n° 1, pp. 158-168. [En ligne]. Disponible à : http://thesai.org/Downloads/Volume8No1/Paper_22-A_Novel_Semantically_Time_Referrer_based_Approach.pdf (Page consultée le 10 août 2018).

KRISHNAGANDHI, Geetha ; DHAS, Suresh Gnana (2016). Web log mining: a study. In *IIOAB Journal*, vol. 7, n° 9, pp. 6-15. [En ligne]. Disponible à : https://www.iioab.org/articles/IIOABJ_7.9_6-15.pdf (Page consultée le 13 août 2018).

MARTINEZ-COMECHE, Juan-Antonio (2017). Determinación de grupos de usuarios de bibliotecas digitales mediante el análisis de ficheros log. In *Revista Española de Documentación Científica*, vol. 40, n° 3, pp. 1-19. [En ligne]. Disponible à : http://dx.doi.org/10.3989/redc.2017.3.1420 (Page consultée le 13 août 2018).

ORSTOM (1955) *Office de la Recherche Scientifique et Technique Outre-Mer : organisation - activités : 1944-1955*. Paris : ORSTOM, 1955, 182 p. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers12-03/010027861.pdf (Page consultée le 10 août 2018).

ROSSI, Pier Luigi (1992). Servers and online bibliographic databases in developing countries: the African reality. In: *Raitt D.I. (ed.). Online information 92*. Oxford : Learned Information, pp. 431-435. ISBN 0-904933-83-0. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-

doc/pleins_textes/pleins_textes_6/b_fdi_35-36/41308.pdf (Page consultée le 10 août 2018).

ROSSI, Pier Luigi (1997) Economie et portabilité : une chaîne d'édition électronique destinée à la dissémination de l'information primaire. In : *Forum initiatives 97*. Hanoi : AUF, 6 p. multigr. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_6/divers1/010022348.pdf (Page consultée le 10 août 2018).

ROSSI, Pier Luigi ; Ngoma-Mouaya Marcel (2000). "Pleins_Textes" : IRD (Institut de Recherche pour le Développement) electronic library. In: *Online information 2000 proceedings*. Oxford : Learned Information, pp. 201-206. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_5/TAP/010024168.pdf (Page consultée le 10 août 2018).

ROSSI, Pier Luigi ; THIAW, Anastasie (2012) Log analysis and text mining on internet access to dissertations of the INSEPS (Institut National Supérieur de l'Education Populaire et du Sport) Dakar, Sénégal. In *African Research and Documentation*, vol 118, pp. 79-90. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers13-05/010058664.pdf (Page consultée le 10 août 2018).

ROSSI, Pier Luigi ; TRAORE Minata ; MAÏGA DIALLO, Fatoumata (2018) Publications en libre accès des universités du Burkina Faso : analyse d'impact et visibilité internationale. In *027.7 Zeitschrift für Bibliothekskultur*, vol. 5, n° 1, pp. 52-64. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers18-02/010072183.pdf (Page consultée le 10 août 2018).

ROUX-FOUILLET, Jean-Paul (1988) Horizon : base bibliographique ORSTOM : présentation. In : *Séchet Patrick (ed.). Séminfor 1, premier séminaire informatique de l'ORSTOM : bases de données et systèmes d'information : quelles méthodes ?* Paris: ORSTOM, pp. 285-296. [En ligne]. Disponible à : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_4/colloques/26249.pdf (Page consultée le 10 août 2018).

SHNEIDERMAN, Ben (1992). Tree visualization with tree-maps: 2-d space-filling approach. In *ACM Transactions on Graphics*, vol. 11, n° 1, pp. 92-99. doi.org/10.1145/102377.115768.

 Dépôt Légal : 2018MO4831 - ISBN : 978-9920-36-568-0

Rossi Pier Luigi. Indicateurs de consultation, indicateurs stratégiques : leur production à partir de l'analyse des consultations des documents d'une archive en libre accès .
In : Ibnlkhayat N. (ed.), Bachr A.A. (ed.), Benchakroun A. (ed.), Roudiès O. (ed.) Le libre accès à la science : fondements, enjeux et dynamiques : actes du 3e colloque international sur l'Open Access. Rabat : Centre National de Documentation, 2018, (1), p. 222-237. (Informaton and Communication Sciences e-Book Series ; 1).
International Colloquium on Open Access (ICOA) : Open access to Science : Foundations, issues and dynamics, 3., Rabat (MAR), 2018/11/28-30. ISBN 9789920365680

http://www.documentation.ird.fr/hor/fdi:010074292