FREE ACCESS TO SCIENTIFIC PUBLICATIONS FOR DEVELOPING COUNTRIES: THE RESEARCH ARCHIVE OF THE FRENCH NATIONAL RESEARCH INSTITUTE FOR SUSTAINABLE DEVELOPMENT (IRD)

By Pier Liugi Rossi
rossi@ird.fr
Institut de recherche pour le développement (IRD), Bondy, France

Introduction

The IRD (French National Research Institute for Sustainable Development)[1] is a French research institute serving the Mediterranean and intertropical countries that make science and innovation one of the first levers of their development. It is a French public science and technology establishment (EPST) which is under the dual supervision of the Ministry of Higher Education and Research and the Ministry of Foreign and European Affairs[2].

The Decree of 1 December 1955 (ORSTOM 1955) organising the services of the Office for Scientific and Technical Research Overseas, published in the Official Journal of the French Republic of 21 December 1955, defines the existence of a documentation centre under the direction of the Institute. Article 6 of the decree states that "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."[3].

Since 1955 the Institute has created an institutional archive with the intention of preserving and disseminating its scientific productions. This institutional archive is currently made up of 96,000 documents and constitutes the IRD documentary collection (FDI).

This patrimonial collection was computerised from 1986 with the creation of the Horizon bibliographic database (Roux-Fouillet 1988). As early as 1996 we initiated its digitisation (Rossi 1997) and 65,000 documents are freely available on the internet (Rossi, Ngoma-Mouaya 2000).

Given the establishment of research centres and assignments, throughout the history of the Institute a significant part of the scientific work carried out (more than 38,000 documents, about 60% of the documents in free access) concerns the countries of the African continent.

Digitisation methods and protocols

When we initiated the digitisation of the collection in 1996, we developed specifications for the production of digital files, both for files produced from digitisation and for files produced from digital sources (files from processing software text or desktop publishing software).

Our choice fell on the pdf format because it ensured a portability on several operating systems, efficient management of the objects composing the documents (text, fonts, images...), a size and optimised characteristics for the internet. We set the scanning resolution at 300 dpi with non-destructive black-and-white image compressions and jpeg compression for grayscale and color images. All scanned documents are processed by optical character recognition software, in automatic mode and without correction of any recognition errors. Particular attention is paid to verifying the integrity and quality of the documents obtained.

Log files[4] and data filtering

The website hosting IRD's collection works with Apache HTTP Server[5]. This software stores the history of the consultations concerning the exchanges of files between the server and the client (the machine which consulted the files available on the server). The format of the history file, in "combined" mode[6], contains information (fig.1) including: the IP address of the client (e.g. 200.113.248.148), the time at which the request was received (e.g 03/Nov/2016:01:53:47 +0100), the URL of the consulted file (e.g /exl-doc/.../010038051.pdf), the status code returned by the server (e.g 200), the site from which the client has launched the request[7] (e.g http://www.google.ht ...), the possible question asked by the customer (e.g. importance de la géographie humaine), the browser used by the client (e.g. Safari) and the type of device used e.g. mobile ...).

```
200.113.248.148 - - [03/Nov/2016:01:53:47 +0100] "GET /exl-doc/pleins_textes/divers11-03/010038051.pdf HTTP/1.1" 200 15655 "http://www.google.ht/search?hl=fr&client=ms-android-att-us&source=android-browser-key&q=importance+de+la+géographie+humaine &gws_rd=cr&ei=d4oaWLjYCsTGmQGV1Z0w" "Mozilla/5.0 (Linux; U; Android 2.1-update1; fr-be; SonyEricssonX10a Build/2.1.A.0.492) AppleWebKit/530.17 (KHTML, like Gecko) Version/4.0 Mobile Safari/530.17"
```

Fig. 1: Example of a transaction line of the Apache server log file, combined format

To carry out our study, we geolocated the IP addresses using the GeoLite Country database of MaxMind[8]. In the example of fig 1 the IP address is located in Haiti.

The log file is filtered conserving only lines that have access to pdf files with status "200"[9]. The resulting file is then broken down into several sub-files. The field that contains the browser type also includes the signature of the search robots. This makes it possible to generate log files relating to indexing that by the search engines, in particular by the Google search engine (fig. 2).

```
66.249.64.158 - - [01/Nov/2016:05:56:05 +0100] "GET /exl-
doc/pleins_textes/fan/010011811.pdf HTTP/1.1" 200 251739 "-"
"Googlebot/2.1 (+http://www.google.com/bot.html)"
```

Fig. 2. Example of a transaction line concerning the indexing of a pdf file by the Google robot (Googlebot / 2.1)

Then we discard and generate a “spam” file with all the lines of the log file which mask by a dash ("-") the referer field (the field of the site from which the client has launched a request) or the navigator field: we presuppose that queries that hide these two fields can be assimilated to DoS attack[10] or spam referer[11]. According to the same logic, we also discard the accesses coming from the same IP address to consult the same file which are realised within less than five minutes.

The filtered job file that contains all server accesses contains 2,363,329 rows, which is 39% of the original raw file. The robots file is 42% and the spam file is 19% of the original raw file.

Geographical accesses distribution

After these different stages of filtering (Kaur, Aggarwal 2017), we have a file that consists (mostly) of "human access". It is ready to produce statistics and accesses indicators (Rossi P.L., Thiaw A. 2012 & Rossi P.L., Traore M., Diallo F.M. 2017).

Since all the accesses have been geolocated, it is possible to produce accesses statistics according to the geographical distribution by countries and by continents. We present these results using "treemap" graphs[12] that are suitable for presenting this type of data (fig. 3).

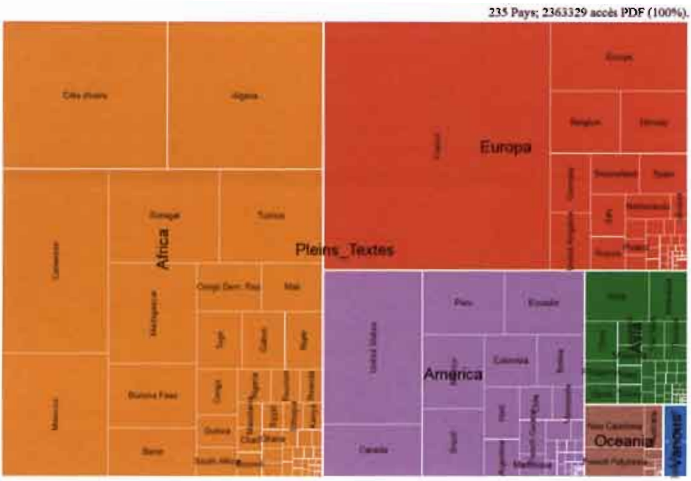


Figure 3

The accesses are located in 235 countries including 55 African countries. France is the first country where consultations are located with 427,960 accesses, 18% of the total. For the countries of the African continent there are 1,107,007 accesses, 47% of the total (Europe: 29%, America: 17%, Asia: 4.4%, Oceania: 1.8%). Côte d'Ivoire is the first country in Africa with 183,725 accesses, 7.8% of the total. The Maghreb countries (Algeria, Morocco and Tunisia) generate 345,717 accesses, 15% of the total.

South Africa is the leading English-speaking country with 9,117 accesses, 0.4% of the total. The United States is the leading country in the Americas with 116,231, 4.9% of the total. India is the leading country in Asia with 24,723 accesses, 1% of the total. New Caledonia is the first country in Oceania with 19,293 accesses, 0.8% of the total.

These statistics can also testify to tense geopolitical contexts characterising some African countries. Our Institute has, for example, produced a large number of publications on Chad (1,344 documents) and Central African Republic (787 documents) which are freely accessible. The number of consultations in these countries is relatively low: for Chad, in 2016 there were 4,611 accesses, i.e. 0.20% of the total and for the Central African Republic 2,220 accesses, 0.09% of the total. These results can be explained by taking into account the number of internet users in relation to the country's population (internet penetration), which are respectively 5% for Chad and 5.4% for Central African Republic (Internet World Stats 2018). These factual elements are amplified by the severe political and social instability that has characterised these countries for several years. The price of internet access is also a limiting factor in several African countries (Cable.co.uk 2017).

The accesses geolocation analysis can be cross-referenced with the language of the documents. There are 1,851,928 accesses to document in French, located in 222 countries including 55 African countries (fig. 4). France is the foremost country where consultations are located with 404,939 accesses, 22% of the total. For the countries of the African continent there are 1,034,657 accesses, 56% of the total (Europe: 32%, America: 8%, Asia: 1.2%, Oceania: 1.9%). Côte d'Ivoire is the leading country in Africa with 178,616 accesses, 9.6% of the total. The Maghreb countries (Algeria, Morocco and Tunisia) generate 327,957 accesses, 18% of the total. This data filtering (language of document: French), makes Haiti appear in the Americas and Lebanon in Asia.

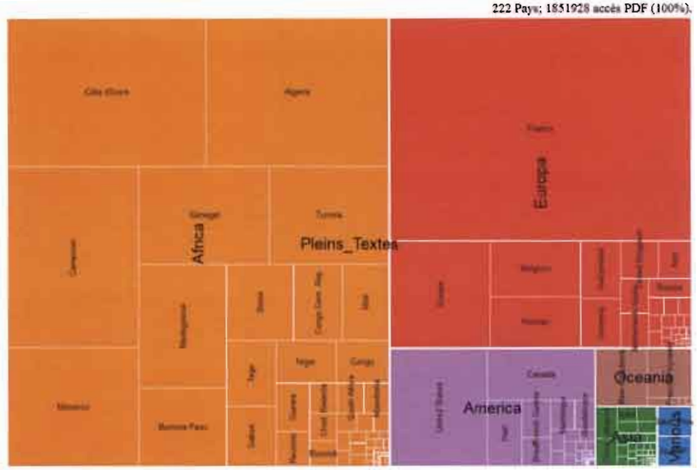


Figure 4

There are 233,775 accesses to documents in English, located in 229 countries including 54 African countries (fig. 5). Asia is the foremost continent where accesses are located with 72,596 accesses, 31% of the total. The United States is the leading country where accesses are located with 35,024 accesses, 15% of the total. For the countries of the African continent there are 36,752 accesses, 16% of the total (Europe: 23%, America: 27%, Oceania: 3.4%). Nigeria is the first country in Africa with 7.171 accesses, i.e. 3.1% of the total. This data filtering (language of document: English) shows the United Kingdom (3.5% of the total) next to France (3.9% of the total) and highlights New Zealand (0.5% of the total), Fiji (0.23 % of the total) and Vanuatu (0.15% of the total) in Oceania.

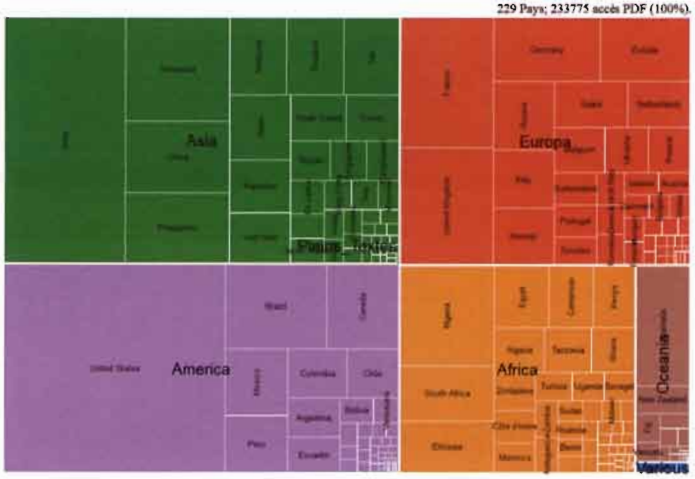
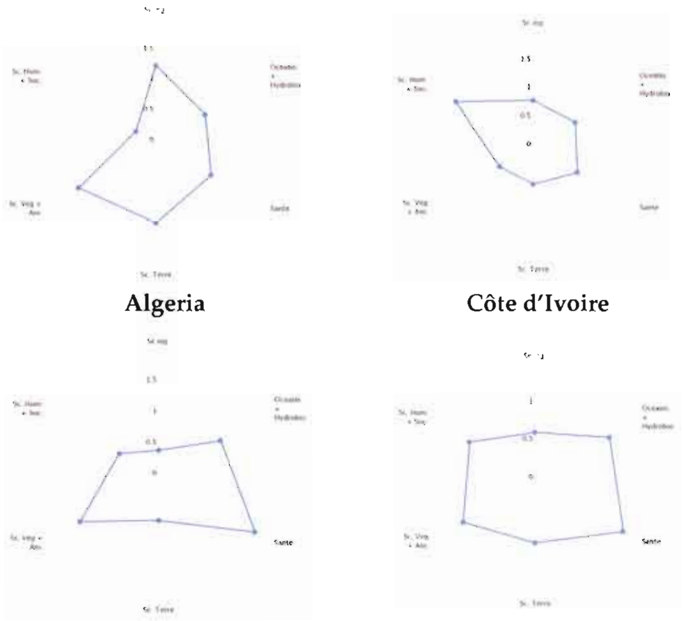


Figure 5

Themes and languages of accessed documents

In the bibliographic database of IRD collection each document is associated with a theme (indexing treatments carried out by the IRD documentalists). The bibliographic database themes are: Engineering Sciences (Sc. Ing.), Oceanography and Hydrobiology (Oceano + Hydrobio.), Health (Sante), Earth Sciences (Sc. Terre), Plant and Animal Sciences (Sc. Veg + Ani), Humanities and Social Sciences (Sc. Hum. + Soc.).

The access data can be filtered in relation to a country and associated with the themes of the documents. These "country" data can be compared to the overall distribution of accesses by theme. This shows an index of interest to the themes in relation to each country (fig. 6).



Algeria **Côte d'Ivoire**

Kenya **South Africa**

Fig. 6: Index of interest to the themes in relation to for African countries

For Algeria, documents concerning Engineering Sciences, Earth Sciences and Plant and Animal Sciences are over accessed. Documents concerning Humanities and Social Sciences are under accessed.

For Côte d'Ivoire documents concerning Humanities and Social Sciences are over accessed and all the other topics are under accessed.

For Kenya documents concerning Oceanography and Hydrobiology and Health are over accessed. Documents concerning Engineering Sciences, Earth Sciences and Humanities and Social Sciences are under accessed.

For South Africa documents concerning Oceanography and Hydrobiology and Health are over accessed. Documents concerning Engineering Sciences and Earth Sciences are under accessed.

As regards the language of the accessed documents (fig. 7), most of the documents consulted from the French-speaking countries (Algeria, Côte d'Ivoire) are in French. On the other hand, for the English-speaking countries the accesses distribution in relation to the language of the documents is much more balanced, with a preference for French in South Africa.

	French	English	Other
Algeria	98.9%	1.1%	–
Côte d'Ivoire	99.3%	0.7%	–
Kenya	44.5%	55.4%	0.1%
South Africa	55.0%	44.6%	0.4%

Fig. 7: Language of the accessed documents

Pedology and hydrology

During 2016 and 2017 we conducted a digitisation campaign for all our scientific productions concerning Pedology (soil science) (Fargier 2015) and Hydrology[13]. These are two disciplines that have characterised the scientific history of our institute and have produced a substantial number of documents particularly significant for the countries in which they have been made: both in Africa and in Latin America.

The documents' accesses concerning Pedology are number 283,859, located in 205 countries including 54 African countries. These accesses concern 6,698 documents which have therefore been consulted, on average, 42 times for the year 2016. France is the first country where consultations are located - with 49,998 accesses, 18% of the total. For the countries of the African continent there are 157,729 accesses, 56% of the total (Europe: 27%, America: 13%, Asia: 2.9%, Oceania: 0.7%). Algeria is the first country in Africa with 36,168 accesses, 13% of the total. The Maghreb countries (Algeria, Morocco and Tunisia) generate 75,503 accesses, 27% of the total. Three documents, concerning soil analysis techniques, each count more than 4,000 accesses each for the year 2016[14].

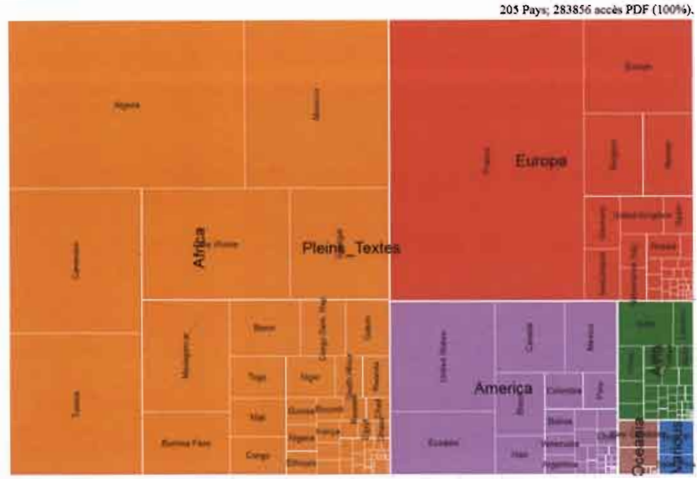


Figure 8

Hydrology documents generate 169,762 accesses located in 202 countries, including 55 African countries. These accesses concern 5,302 documents which were thus consulted, on average, 32 times for the year 2016. France is the first country where consultations are located with 29,614 accesses, 17% of the total. For the countries of the African continent there are 85,184 accesses, 50% of the total (Europe: 26%, America: 20%, Asia: 2.5%, Oceania: 0.9%). Algeria is the first country in Africa with 16,849 accesses, 9.9% of the total. The Maghreb countries (Algeria, Morocco and Tunisia) generate 41,039 accesses, 24% of the total. Brazil is the first country in the Americas with 8,541 accesses, 5% of the total. Two documents concerning hydrological analysis techniques and the construction of dams count more than 4,000 accesses each for 2016[15].

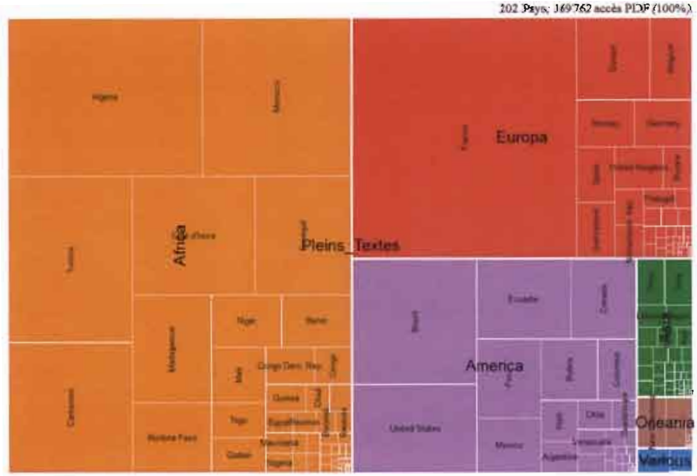


Figure 9

Conclusion

The project that we launched in 1996 enabled the digitisation and open access of about 66% of the IRD's scientific output (more than 65,000 documents in pdf format). The project continues with an annual increase of about 4,000 pdf files.

The analysis of accesses statistics shows the strong impact of the documents on the scientific contexts of the African countries. The combination of these data with the themes and languages of the documents shows specificities for each country (thematic) or for the linguistic areas (French-speaking countries versus English-speaking countries) of the African continent.

The analysis illustrates that the most consulted documents are often documents published in the '1990s dealing with methodologies, analysis techniques and construction of civil engineering structures (dams).

References

Cable.co.uk (2017). Study of broadband pricing in 196 countries reveals vast global disparities in the cost of getting online www.cable.co.uk/media-centre/release/new-worldwide-broadband-price-league-unveiled/

Fargier N. (2015) Numériser la littérature grise scientifique. I2D Information, données et documents, 52(1), p. 61-62 www.cairn.info/revue-i2d-information-donnees-et-documents-2015-1-page-61.htm

Internet World Stats (2018). Internet Users users Statistics statistics for Africa. www.internetworldstats.com/stats1.htm

ITU/UNESCO Broadband Commission for Sustainable Development (2017). The state of broadband 2017: broadband catalyzing sustainable development. ITU, Unesco, 104 p. www.itu.int/dms_pub/itu-s/opb/pol/S-POL-BROADBAND.18-2017-PDF-E.pdf

Kaur, N, Aggarwal, H (2017) A Novel semantically-time-referrer based approach of web usage Mining for Improved Sessionization in Pre-Processing of Web Log. International journal of advanced computer science and applications 8(1), p. 158-168. thesai.org/Downloads/Volume8No1/Paper_22-A_Novel_Semantically_Time_Referrer_based_Approach.pdf

ORSTOM (1955) Office de la Recherche Scientifique et Technique Outre-Mer : organisation - activités : 1944-1955. Paris, 182 p. horizon.documentation.ird.fr/exl-doc/pleins_textes/divers12-03/010027861.pdf

Rossi P.L. (1992) Servers and online bibliographic databases in developing countries : the African reality. In : Raitt D.I. (ed.). Online information 92. Oxford : Learned Information, p. 431-435. ISBN 0-904933-83-0. horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_6/b_fdi_35-36/41308.pdf

Rossi P.L. (1997) Economie et portabilité : une chaîne d'édition électronique destinée à la dissémination de l'information primaire. In : Forum initiatives 97. Hanoi 25-26 octobre 1997, 6 p. multigr. horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_6/divers1/010022348.pdf

Rossi P.L., Ngoma-Mouaya M. (2000). "Pleins_Textestextes" : IRD (Institut de Recherche pour le Développement) electronic library. In: Online information 2000 proceedings. Oxford : Learned Information, p. 201-206. horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_5/TAP/010024168.pdf

Rossi P.L., Thiaw A. (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. African Research and Documentation, 118, p. 79-90. ISSN 0305-826X. horizon.documentation.ird.fr/exl-doc/pleins_textes/divers13-05/010058664.pdf

Rossi P.L., Traore M., Diallo F.M. (2017) Publications en libre accès des universités du Burkina Faso : analyse d'impact et visibilité internationale. 027.7 : Zeitschrift für Bibliothekskultur, 5 (1), p. 52-64. ISSN 2296-0597, horizon.documentation.ird.fr/exl-doc/pleins_textes/divers18-02/010072183.pdf

Roux-Fouillet J.P. (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, p. 285-296. ISSN 0767-2896. horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_4/colloques/26249.pdf

Shneiderman B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on Graphics, v. 11-1, p. 92-99. doi:10.1145/102377.115768

Notes

[1] See: en.ird.fr/the-ird/presentation Accessed: 08/03/2018

[2] IRD was created in 1943 under the name of Office of Colonial Scientific Research (ORSC). In 1953 it became Orstom (Office of Scientific and Technical Research Overseas). In 1984 it took the name of French Institute of scientific and technical research for development and cooperation, while retaining its acronym and became IRD in 1998. See: en.ird.fr/ird.fr/the-ird/history. Accessed: 08/03/2018.

[3] "The Documentation Center is responsible for the processing, preservation and dissemination of scientific and technical documentation relating to the activities of O.R.S.T.O.M."

[4] See the definition of computer history, log and log at: en.wikipedia.org/wiki/Log_file

[5] See: httpd.apache.org

[6] See: httpd.apache.org/docs/2.4/en/logs.html

[7] It is the referer field. See: en.wikipedia.org/wiki/HTTP_referer

[8] GeoLite country and GeoLite city free versions are freely available at: dev.maxmind.com/geoip/legacy/geolite

[9] See: tools.ietf.org/html/rfc2616#page-58. The status "200" indicates the success of the request. With the GET method an entity corresponding to the requested resource is sent with the response.

[10] See: en.wikipedia.org/wiki/Denial-of-service_attack

[11] See: en.wikipedia.org/wiki/Referrer_spam

[12] See: en.wikipedia.org/wiki/Treemapping. See also: "Treemaps for Space-Constrained Visualization of Hierarchies: The History of Treemap Research at the University of Maryland. Started Dec. 26th, 1998 by Ben Shneiderman". www.cs.umd.edu/hcil/treemap-history/index.shtml

[13] Many documents concerning these disciplines had already been digitized since 1996. However, for a large number of documents containing thematic maps and plans, digitization had been posticipated, given their complexity and therefore higher costs for their digitization.

[14] These are : **Pétard, Jean (1993)** *Les méthodes d'analyse : tome 1. Analyse de sols* ; **Yoro, G. - Godo, G. (1989)** *Les méthodes de mesure de la densité apparente : analyse de la dispersion des résultats dans un horizon donné* ; **Aubert, Georges (1962)** *Cours de pédologie générale. Processus de formation des sols : profils de sols ferrallitiques.* It should be noted that these highly consulted documents have not been published recently. These three documents are all widely consulted by Internet users located in the Maghreb.

[15] These are: **Dubreuil, Pierre (1974)** *Initiation à l'analyse hydrologique (dix exercices suivis des corrigés)* ; **Molle, François - Cadier, Eric (1992)** *Manual do pequeno açude.* It should be noted that these documents have not been published recently and that the second document generates 4,358 accesses from Brazil (90% of accesses of this document).