LOG ANALYSIS AND TEXT MINING ON INTERNET ACCESS TO DISSERTATIONS OF THE INSEPS (INSTITUT NATIONAL SUPERIEUR DE L'EDUCATION POPULAIRE ET DU SPORT) DAKAR, SENEGAL

By Pier Luigi Rossi[1] and Anastasie Thiaw[2]

Introduction

The dissertation collection of INSEPS (Higher National Institute of Popular Education and Sport, Dakar, Senegal) consists of 152 documents (PDF format) related to academic work submitted between 2005 and 2008 (2005: 21, 2006: 29, 2007: 45 and 2008: 57) as well as all references to available dissertations of INSEPS' library (imported from a CDS/ISIS[3] database).

These have, since January 2011, been hosted on the BEEP (Electronic libraries in partnership) website[4] which uses the Greenstone software[5] (Rossi, 2011). Pdf files of the collection were created either by scanning paper or by converting electronic versions (Word files). The collection of electronic documents was achieved as part of the SIST[6] (System for Scientific and Technical Information) project funded by the MAEE[7] (French Ministry of Foreign and European Affairs).

Through this study, an attempt is made to better define the audience of documents of the collection, following their setting on the web. An attempt is also made to measure the volume and change of this audience over time. The variety of users is examined (geographical origin, concentration or dispersion vis-à-vis the entire funds). Finally, the users' various concerns are studied. The investigation is based on analysis of modes and frequencies of internet consultations of documents in PDF format. This approach is based on the use of log files of the BEEP Apache[8] server.

Methodology

The access log of an Apache server is used to record all the transactions to access files hosted by the server and consulted by users[9]. The format of the BEEP access log is a "combined log format". It helps to know, in addition to the "standard log format" information, the header "Referer"[10] and the "User-Agent" of the request.

To analyse INSEPS' dissertation consultations (pdf files of each dissertation), a file of "effective access" was created through several filtering steps on the lines of the Apache log file:

(1) selection of lines relating to pdf files of INSEPS' collection,
(2) selection of access lines with a status of "200"[11],
(3) exclusion of spiders access lines,
(4) exclusion of "HEAD" method access lines[12],
(5) exclusion of spam access lines[13].

The IP address of each line of the "effective consultation" file is then identified by country. Ip address resolution is done using a specific php script that includes the "MaxMind GeoIP Country Database"[14].

Results

(1) The audience volume: Evolution along time.

Analysis of "effective consultations" of the documents of the INSEPS collection shows the distribution of the consultations per month (Table 1) and the average smoothed consultations per quarter (Table 2).

After the first two months, the number of consultations quickly rose to a "cruising level", peaking at 2,515 consultations per month in May 2011 (a monthly average of 16.5 consultations per document). March 2012 recorded the highest number of consultations (2701, with a monthly average of 17.7 per document). Compared to March 2011 the increase was 66.8 per cent.

The average of consultations per month smoothed by quarter (Table 2) also shows the rapid rise in the average number of consultations after the first months of collection implementation on the internet. The first quarter of 2012 saw the highest monthly average (smoothed for the first three months): 15.7.

A calendar-specific feature appears. Months of school holidays (July-September 2011) saw a decline in consultations with a "trough" in August 2011 (the monthly average is 10 consultations per document). In contrast, the typical time of the preparation of mid-term or final examinations (March-May) gives rise to a peak of consultations.

This rate suggests that a substantial number of the readership has academic origin, probably students tasked to produce dissertations similar to those

Table 1: Distribution of consultations per month

Jan-11	Feb-11	Mar-11	Apr-11	May-11	Jun-11	Jul-11	Aug-11	Sep-11	Oct-11	Nov-11	Dec-11	Jan-12	Feb-12	Mar-12
905	1322	1619	1733	2515	2272	2030	1529	1621	2205	2208	2007	2066	2390	2701

Table 2: Average file consultations smoothed by quarter

Jan-Mar 11	Apr-Jun 11	Jul-Sep 11	Oct-Dec 11	Jan-Mar 12
8.4	14.3	11.4	14.1	15.7

presented here. But, with this lack of direct data on individuals who consult documents, it is necessary to remain cautious.

(2) Frequency of consultations per document: Concentration, relative dispersion.

A second approach involves the statistical distribution of consultations on the various documents of the collection. Following the Pareto principle[15], it could be assumed that only 20 per cent of the documents accounted for 80 per cent of consultations.

To compare this principle with the distribution of consultations of documents from the INSEPS collection account is taken of consultations during the first quarter of 2012 (Figure 1).

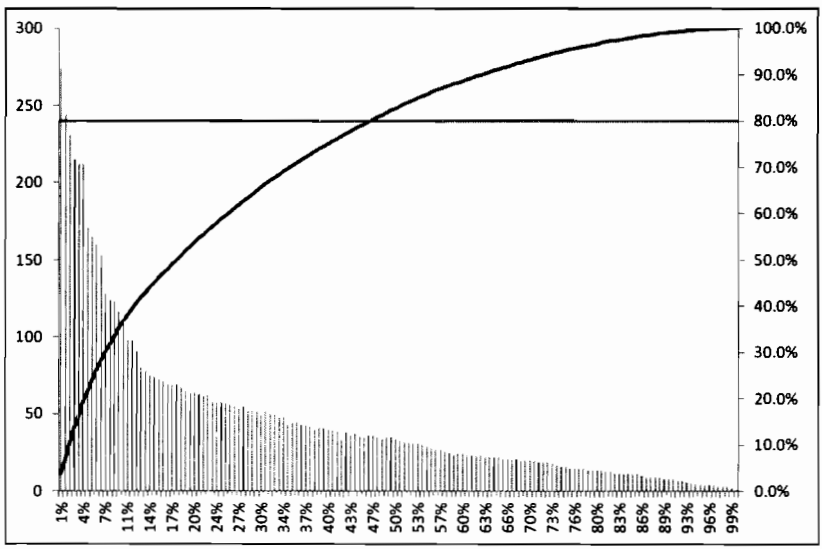

Figure 1: Frequency dissertation accesses (first quarter 2012).

It is observed that the distribution of consultations is not focused around 20 per cent of the most consulted documents. Indeed, the 80 per cent threshold is reached with 46 per cent of the most consulted documents. Considering 20 per cent of the most consulted documents, only 53.8 per cent of total consultations are reached.

According to these data, an average of 12 per cent of documents are visited more than 30 times per month, 66 per cent of documents are visited 5-30 times per month and 22 per cent of documents are consulted less than 5 times per month.

It can therefore be estimated that:

(1) Publishing data on the internet brings to the collection a remarkable readership, certainly higher than the same "paper documents" available only in a single library,

(2) Publishing data on internet attracts a variety of unexpected interest which reduces the usually experienced concentration phenomenon.

(3) The variety of publics: Geographical origins of consultations.

The preceding remarks can be amplified by studying the geographical origins of the consultations. Figure 2 illustrates the distribution of consultations of Inseps' dissertations per month from a country[16]. The ranking is based on the consultations frequency of the last month (March 2012), and only the first 34 countries are presented.

	Dec 2010	Jan 2011	Feb 2011	Mar 2011	Apr 2011	May 2011	Jun 2011	Jul 2011	Aug 2011	Sep 2011	Oct 2011	Nov 2011	Dec 2011	Jan 2012	Feb 2012	Mar 2012
France	67	353	586	677	681	[illegible]	737	881	575	722	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Senegal	21	109	109	149	170	239	215	233	225	100	149	177	182	164	234	261
Algeria	31	111	148	227	172	301	261	152	161	203	188	158	180	178	219	234
Tunisia	22	65	53	106	240	272	189	110	71	81	125	116	101	98	173	185
United States	3	8	16	13	31	41	58	39	32	49	90	116	101	98	84	124
Morocco	14	42	79	55	60	142	122	116	65	60	105	108	114	103	100	118
Canada	18	42	59	67	76	74	57	32	31	23	36	52	25	36	60	74
Belgium	0	41	72	66	47	106	69	46	74	45	53	57	64	48	47	63
Côte d'Ivoire	2	15	19	18	4	23	29	22	18	6	10	21	19	30	49	54
Benin	2	2	26	35	15	21	30	22	23	8	17	16	5	6	30	44
Cameroon	4	13	27	44	38	40	49	24	21	20	30	20	52	46	50	38
Gabon	1	6	8	6	11	12	5	4	0	1	14	20	19	14	8	27
Switzerland	7	17	21	21	17	26	18	17	20	24	12	13	24	23	19	20
Japan	1	9	8	12	23	20	18	12	8	6	8	14	13	16	13	17
Burkina Faso	1	4	9	9	8	15	9	15	9	3	5	20	12	27	17	16
Italy	1	6	5	6	3	13	5	16	11	9	9	11	10	11	6	11
Niger	0	2	3	3	5	6	2	3	8	0	0	3	8	1	3	10
United Kingdom	3	2	5	1	5	3	6	3	4	9	7	7	5	2	6	9
Spain	1	2	7	7	10	7	9	7	9	1	8	9	15	10	9	8
Lebanon	0	1	4	8	4	9	7	6	2	1	2	6	4	1	3	8
Norway	0	0	1	2	0	0	1	2	0	0	0	0	0	0	0	8
Djibouti	0	2	1	2	10	4	3	1	2	2	2	3	0	0	2	6
Mali	1	0	8	6	10	7	9	4	5	6	5	5	9	1	0	6
Mauritania	0	2	0	1	3	2	1	1	2	2	0	0	2	3	0	6
Togo	0	1	2	2	14	63	72	2	12	7	10	8	1	2	6	6
Germany	1	2	5	2	2	51	175	420	65	173	85	6	2	3	5	5
Madagascar	0	0	5	17	5	7	28	16	4	1	9	3	9	20	10	4
Nigeria	0	0	1	0	1	0	1	2	0	1	0	2	0	3	2	4
Chad	0	0	0	1	0	0	0	1	1	2	1	0	0	0	0	4
Satel. Prov.	0	2	4	4	4	9	12	3	9	2	5	3	5	3	5	3
Burundi	0	1	1	2	0	1	2	1	0	0	0	1	1	0	4	3
Congo	0	4	0	3	0	0	1	5	0	0	2	1	2	6	8	3
Austria	0	5	0	0	2	0	0	1	0	0	0	0	0	2	1	2
India	0	0	1	0	2	0	0	0	0	0	0	0	0	1	0	2

Figure 2: Frequency dissertation accesses per month from a country

France is the country with the largest number of consultations (1287 in March 2012). It exceeds by about five times the number of consultations in Senegal (261 in March 2012) or in Algeria (234 in March 2012). If the Maghreb countries (Algeria, Morocco and Tunisia), are considered together, a significant number of consultations are made in this region (527 in March 2012).

There is a significant increase in consultations in France with effect from October 2011 and peaks in Germany around July 2011. This trend for Germany may be due to the presentation of the BEEP project at the ECAS4 symposium[17], where several German colleagues were present.

The fact that France accounts for a larger number of consultations than other countries, especially Senegal (the country where the dissertations were defended), could be partly explained, taking into account the number of French students in the sports disciplines (7274 STAPS masters level students for the year 2010-2011)[18], compared to Senegal (about 50 final year dissertations per year).

It is also noted that most of the consultations are made in francophone countries: the dissertations are in French and the effect of language on the origin of consultations here is undeniable. The major exception is the United States, fifth score of consultations, before Canada and many African countries. As mentioned above, another exception is Germany, with a peak in the summer of 2011.

The distribution of the entire country consultations of the collection (Figure 2) is the result of consultations for each document. However, the analysis of consultations for each document shows more specific distributions.

Figure 3 shows the distribution by country of the collection's most viewed document. For this dissertation, concerning speed improvement in Senegalese football, the consultations account for Maghreb countries is 44 per cent (selection of the three countries: darker area) and France 38 per cent.

For the document that is the fifth most consulted dissertation (Figure 4), the specific place of a Senegalese location (Thiès) for a document concerning physical education in the elementary school, is "decisive"[19]: Senegal accounts for 31 per cent of the consultations, France 17 per cent and the Maghreb 12 per cent.

These two examples suggest that the geographical location of consultations for each document has a different distribution, i.e. very different from the distribution calculated for all documents. It should therefore be borne in mind that the general trend cannot be extrapolated from each specific document.

(4) The routing of requests: Diversity of publics and their concerns.
The data recorded in the log files provide information on the questions asked by users that can lead to pdf documents of the collection.

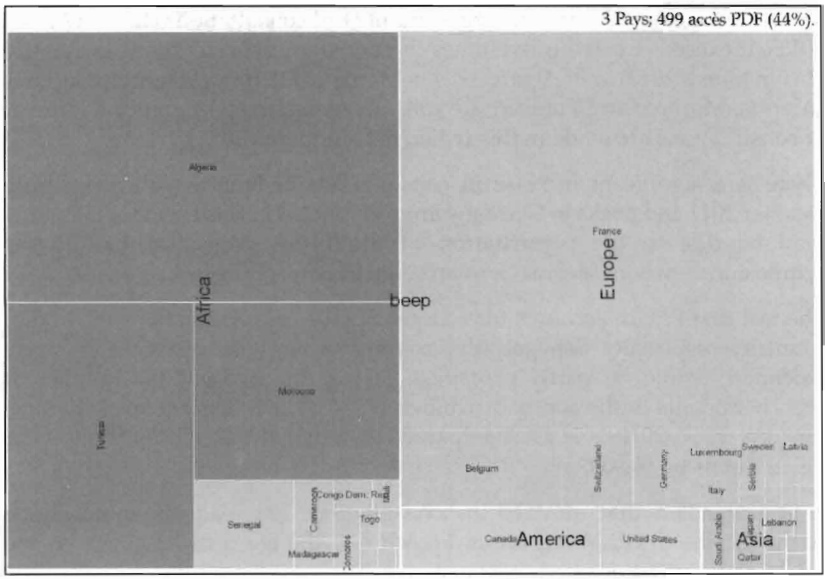

Figure 3: Distribution by country of the document most viewed

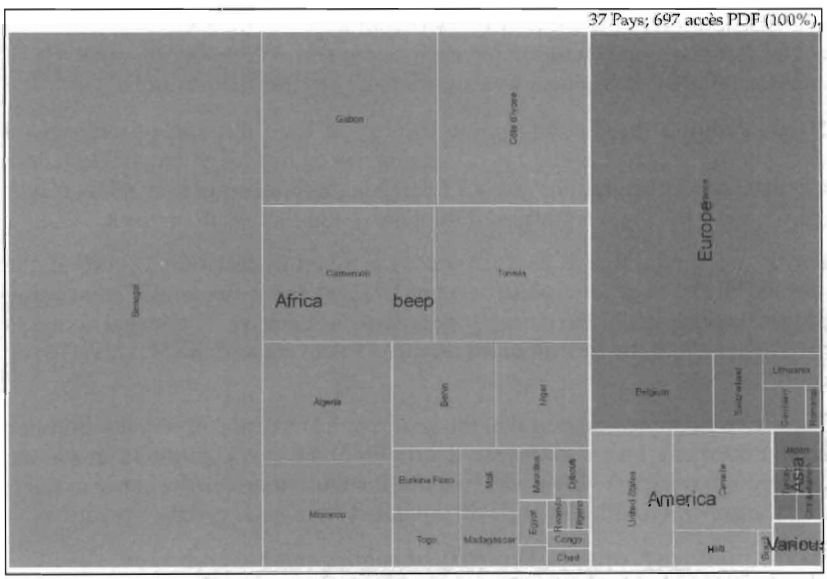

Figure 4: Distribution by country of a document concerning Thiès (Senegal).

With a programme written in php, extraction and treatment are made of the text of all questions that led to the consultation of document during the first quarter of 2012. The text of the questions is analysed with TreeTagger[20], a language independent part-of-speech tagger. Given that the dissertations are written in French, most of the questions are in French and therefore French settings of TreeTagger were applied. After the treatment, nouns, adjectives and verbs were retained to be used as "simplified questions"[21]. Our programme calculates the challenges of these grammatical elements in each "simplified question" and produces a file that describes their relationships and their frequencies.

At first, this approach allows identification of the most frequent query items. "Keywords" and "key associations" are identified. The most common of these indicate the concerns of a public of specialists ("physical quality", "fitness" and "physical preparation"). This public is surely retained by the type of documents presented here, which concern practices and field observations rather than theories or a sophisticated scientific praxis.

The approach also allows identification of more specific topics of interest (different types of sports, such as "traditional wrestling", which is the object of passion, especially in Senegal) or more original ("woman's veil and sport": a question of ethical, ideological and practical connotation with a significant recurrence).

The graph of co-concurrences of these grammatical elements (nouns, adjectives and verbs) in each "simplified question" is generated with Gephi[22], an open-source software for visualising and analysing large network graphs. Each element of "simplified question" is represented by a vertex.

It was made accessible on the web[23] using gexf-js[24] (Figure 5). The graph can thus be viewed and analysed by many users, and provides a dynamic and selective visualisation (integrated search tool).

The selection of a word of the graph allows viewing of related words as a graph and as a list (left side of the window), which also indicates the frequency of the links.

This graph has been used to identify the major sports (or sporting activities) in the questions asked by users, leading to consultation of a document of the INSEPS collection. Eight sports were identified: football, wrestling, basketball, taekwondo, jumping, volleyball, handball and karate. A search was made for these terms (or the root of these words: foot, basket, volley, hand) in the full text of the questions (without performing linguistic processing) to calculate their overall frequency and their geographical location (inferred from IP addresses).

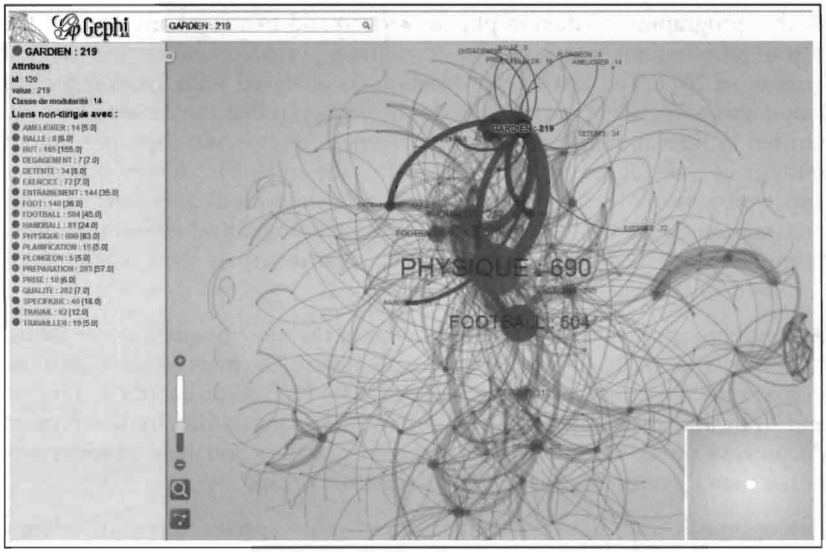

Figure 5: Co-occurrences graph of nouns, adjectives and verbs in "simplified question".

For the geographical location, three entities were identified: France, Senegal and the Maghreb (Algeria, Morocco and Tunisia).

With these data, calculation was made of the **specialisation index** of these geographical entities in relation to sporting activities (Table 3). This index represents the ratio between the frequency of search for an activity from a country vis-à-vis the total searches for the same sport divided by the ratio of the total frequency of searches from a country vis-à-vis the total searches[25].

Table 3: Questions specialisation index by sports and by country

Sports	Senegal	France	Maghreb
foot	0.70	0.85	1.67
wrestling	2.37	0.63	0.18
basket	0.59	1.01	1.05
taekwondo	3.74	0.12	0.39
jump	1.26	0.69	1.00
volley	1.00	0.72	0.97
handball	0.89	1.29	0.91
karate	0.61	1.08	0.95

Questions from **Senegal** denote a strong specialisation on **taekwondo** and **wrestling** and a sub-specialisation on basketball, karate and football.

Questions from **France** denote a slight specialisation on **handball**, a strong sub-specialisation on taekwondo and a sub-specialisation on wrestling, jumping and volleyball.

Questions from the **Maghreb** denote specialisation on **football** and a strong sub-specialisation on wrestling and taekwondo.

Conclusions

Using the case study of INSEPS' dissertations, the objective was to analyse the audience drawn to the web by a quite specialised documents collection, supposedly confined to a fairly confidential diffusion and to study the modes and frequencies of consultation, using Apache log files from the server that hosts the collection. This method provided several interesting indicators.

First, the **internet** gives these papers a **wide visibility**. For the first quarter of 2012, the monthly consultation average was 15.7 per document. 78 per cent of the documents were consulted at least 5 times per month (12 per cent more than 30 times per month).

The **dissertation language** strongly influences the location of the consultations: **France and Francophone countries** have the largest number of consultations but the US is also present.

The large number of consultations from **France** is doubtless due to the large number of **undergraduate/masters students** in that country in the field of sports and probably to the relative rarity of this type of documents (detailed and close to the experience on the ground) on the web.

The **geographical distribution** of consultations varies from one document to another. It is often related to the **type of sport** concerned or the "**local roots**" of the subject discussed by the dissertation.

Analysis of questions asked by users, based on the calculation of the co-occurrences of significant words (nouns, adjectives and verbs), gives an overview of the **most frequent queries** (physical quality, physical preparation and football). It also helps to identify more specific themes (traditional wrestling, woman's veil and sport).

Identified were **eight sports** that matter most to users (football, wrestling, basketball, taekwondo, jumping, volleyball, handball and karate) and a "**specialisation index**" was calculated which differentiates the interests of three regions (France, Senegal and the Maghreb).

These data show the surprising extension of the audience received by such

documents and also the variety of publics and interest on the web.

These indications may be useful for teachers to guide INSEPS' dissertation themes based on the interests of users in certain countries or regions.

Awareness of this potential interest in specialised dissertation collections could encourage other producers to digitise and publish their own collections to develop their impact and to improve the understanding of their public.

References

Agosti, M. and Di Nunzio, G.M. (2007) 'Gathering and Mining Information from Web Log Files', in C. Thanos, F. Borri, and L. Candela (Eds.), *Digital Libraries: R&D, LNCS 4877*, pp.104–113, Berlin: Springer-Verlag.

Agosti, M., Crivellar,i F. and Di Nunzio, G.M. (2012) 'Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction', *Data Mining and Knowledge Discovery*, 24, pp.663-696.

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi: an open source software for exploring and manipulating networks', p.2, International AAAI Conference on Weblogs and Social Media. (http://gephi.org/publications/gephi-bastian-feb09.pdf)

Bostock, M. and Heer, J. (2009) 'Protovis: A Graphical Toolkit for Visualization', p. 8, IEEE InfoVis 2009. (http://mbostock.github.com/protovis/protovis.pdf)

Croll, A. and Power, S. (2009) *Complete Web Monitoring*, p.672, Sebastopol: O'Reilly Media.

Ministère de l'éducation nationale, de la jeunesse et de la vie associative, Ministère de l'enseignement supérieur et de la recherche (2011) *Repères et références statistiques sur les enseignements, la formation et la recherche (RERS 2011)*, p.424, Paris: Direction de l'évaluation, de la prospective et de la performance. (http://cache.media.education.gouv.fr/file/2011/01/4/DEPP-RERS-2011_190014.pdf)

Picard-Aitken, M. and Côté, G. (2010) *Bibliometric analysis of aquaculture research at DFO and in Canada: final report*, p.46, Montréal: Science-Metrix. (http://www.sciencemetrix.com/pdf/SM_DFO_Aquaculture_Research.pdf)

Rossi, P.L. (2011) 'Electronic libraries in partnership: BEEP for Africa', *African Research and Documentation*, 115, pp.69-75. (http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-11/010052810.pdf)

Notes

[1] IRD (Institut de recherche pour le développement), Bondy, France. mail: rossi@ird.fr

[2] INSEPS (Institut National Supérieur de l'Education Populaire et du Sport) [2], Dakar, Sénégal. mail: anastasiethiaw@hotmail.com

[3] http://en.wikipedia.org/wiki/CDS_ISIS

[4] http://www.beep.ird.fr

[5] http://www.greenstone.org

[6] http://www.sist-sciencesdev.net

[7] http://www.diplomatie.gouv.fr

[8] http://httpd.apache.org

[9] http://httpd.apache.org/docs/current/logs.html

[10] http://en.wikipedia.org/wiki/HTTP_referer

[11] http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10.2.1

[12] http://www.w3.org/Protocols/rfc2616/rfc2616-sec9.html#sec9.4

[13] Access to PDF documents was calculated by eliminating recursive access to the same pdf file. Two accesses to the same file with the same IP address are taken into account if there is a difference of more than five minutes between the two accesses.

[14] http://www.maxmind.com/app/country

[xvi] http://en.wikipedia.org/wiki/Pareto_principle

[15] This "heatmap" and the "treemap" (figures 3, 4) are made with Protovis (Bostock M., Heer J., 2009).

[16] http://www.nai.uu.se/ecas-4/panels/141-156/panel-145/Pier-Luigi-Rossi-full-paper.pdf

[17] http://cache.media.education.gouv.fr/file/2011/01/4/DEPP-RERS-2011_190014.pdf

[18] The title of document is: 'La problématique de l'enseignement de l'éducation physique à l'école élémentaire : le cas de la commune de Thiès'.

[19] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[20] Sentences that contain only masculine singular nouns and adjectives, as well as infinitive verbs.

[21] http://gephi.org/

[22] http://www.beep.ird.fr/graphe/index.html#inseps.gexf

[23] https://github.com/raphv/gexf-js

[24] **Specialisation index (SI) = (QSc/QS) / (Qc/TQ)**

Where:
QSc = frequency of questions of a sport from a country (e.g. question about "foot" from Senegal).
QS = frequency of questions of a sport (e.g. questions about "foot").
Qc = frequency of questions from a country (e.g. questions from Senegal).
TQ = frequency of all questions.
See : Picard-Aitken M., Côté, G. (2010).