Internet 1.0: early users, early uses

Thomas BEAUVISAGE*, Valérie BEAUDOUIN*,
Houssem ASSADI*

Abstract

We are presenting here a set of tools that enabled us to build a detailed picture of Internet uses, connecting it to users with precisely defined socio-demographic profiles. On the one hand, our work is based on representative panels of the Internet-connected population monitored over time; this allows us to get a sociologically sound vision of Internet uses. On the other, the data we analysed is collected by a software probe installed on users' PCs, thus allowing us to implement a user-centric approach. We firstly justify this methodological choice with respect to other possibilities (especially server-centric approaches). Then, we present the software modules we developed in order to collect usage data and qualify them. Finally, a few examples of our platform applications are presented: uses typology and Web paths according to the topics of visited Web. These results represent a unique insight into Internet uses in the 2000-2002 period.

Key words: Representative panels, cohort, probes, traffic data, Internet usage, segmentation.

TITRE FRANÇAIS

Résumé

Nous présentons un dispositif technique qui nous a permis de construire une connaissance fine des usages d'Internet durant la période 2000-2002 en les reliant à des utilisateurs dont les profils socio-démographiques sont qualifiés. Le dispositif s'appuie d'une part sur la notion de panels représentatifs de la population connectée à Internet en France, faisant l'objet d'un suivi dans le temps et d'autre part sur un système de recueil des traces techniques. Ce choix méthodologique est justifié au regard des autres possibilités. Les modules d'enrichissement des traces techniques qui permettent d'accéder à la notion d'usage sont ensuite explicités. Enfin, quelques exemples d'applications de la plateforme sont présentés : typologie des usages et des parcours sur le Web en fonction des thématiques des sites visités.

Mots clés : Panels représentatifs, cohorte, sondes, données de trafic, usages d'Internet, segmentation.

* France Telecom R&D – 38-40, rue du Général Leclerc, 92794 Issy-les-Moulineaux, France ; {thomas.beauvisage, valerie.beaudouin, houssem.assadi}@orange-ftgroup.com

Contents

I. *Introduction*
II. *Collecting traffic data on representative panels*
III. *From traffic datat to usage dexcriptors*
IV. *Results: usage profile and trajectories*
V. *Perspectives*
References (18 ref.)

I. INTRODUCTION

In September 2005, T. O'Reilly proposed the Web 2.0 paradigm for describing new services, actors and uses of the Internet [16]. These changes announce the emergence of disruptive uses and business models, supported by innovative software technologies. Our research focuses on the previous generation of Internet: in order to understand transformations related to Web 2.0, it seems important to develop a "historical approach" of what might be called *Internet 1.0*,

Internet 1.0 was the first hybrid "do-everything tool" to allow very different types of activities: information seeking and browsing, e-mail, live conversations, music downloading, network gaming, or e-commerce. It was a new way of bringing together two worlds previously separated: inter-personal communication and mass media and leisure. The digitalisation of culture means that inter-personal exchanges and cultural consumption are coming together in the same digital world. Describing Internet usage is therefore an essential component of the overall knowledge of information and communication technology consumption.

The France Telecom R&D human sciences laboratory has developed methods for detailed analysis of fixed and mobile telephone by cross-referencing traffic analysis and surveys qualifying the households and individuals observed [17]. With the appearance and generalisation of the Internet, it was felt necessary to set up new ways of studying network usages which are enriching the range of communication practices while simultaneously transforming them. The objective was therefore to acquire similar skills in the field of Internet usage to those developed for analysing telephone usage. We needed to faithfully describe the reality of Internet practice, extending beyond the limits of the countless online studies which only reach "Internet addicts", quantitative studies based solely on what people stated and interviews which provided a detailed and comprehensive description of practices but were limited by their samples. Drawing on traffic data from representative panels seemed the most suitable approach. Traffic data was used rather than what people said, as the activities allowed by the Internet all take place in the same situation: in front of a screen with a keyboard and mouse, making it difficult to detect and describe the diversity of the user's practices. The analysis of usage information collected from the user's workstation enables *activity behind the screen to be broken down* in order to *reconstruct Internet surfer profiles* based on their practices. Representative panels were used because we required an overview of a practice in all its diversity.

It was in this context that we launched a partnership in the year 2000 with the audience measurement company NetValue for secondary processing of their panel data. This partnership provided very detailed information on Internet users representative of the population connected at home in France in the year 2000. This first partnership was followed by the SensNet project between France Telecom R&D, Nielsen/Netratings (formerly NetValue), LIMSI and Paris University III[1]. The aim of the latter project was to set up a system for semantic categorisation of Web usage paths using more extensive panel data (three years of observation: 2000 to 2002; three countries: France, the United Kingdom and Spain).

In this article we will begin by justifying our choice to work with a representative panel on the basis of technical usage information collected from the user's workstation. We will then describe the different enrichment procedures that were used. Finally, we will present some results: firstly, user typologies based on their Internet practices between 2000 and 2002 and, secondly, a typology exploiting the information on the content visited. The results presented in this paper are limited to the French panel, the only one for which we had the opportunity to achieve a very detailed data analysis process.

II. COLLECTING TRAFFIC DATA ON REPRESENTATIVE PANELS

The system implemented meets a dual requirement: working with representative panels and collecting usage data by means of technical tracking information. In this section, we will justify this methodological choice.

II.1. Panels and cohorts

The activity consisting in measuring Internet usage emerged almost at the same time as the Internet itself, using all the research methods available, including monitoring of trials with early adopters, questionnaires, quantitative and qualitative surveys, etc. Data processing approaches from objective traffic measurements have been rare, and surveys designed to be representative of the population connected to the Internet have been even rarer. In this context, audience measurement companies have played a fundamental role. There were three in France in the year 2000, MMXI, NetValue and Netratings. Following a series of bankruptcies and buyouts, Internet measurement in France is now in the hands of a single player, a joint-venture between Médiamétrie and Nielsen/Netratings.

Audience measurement is a key input for companies with an audience-based business model and for which advertisement revenues are crucial (see [11] and [12]). In this context, audience data must meet 2 requirements, the population must be representative and the browsing data must be comprehensive. The representative nature of the panel was maintained month by month with a telephone survey of a very large sample, assessing the population connected to the Internet and its characteristics. The panel was adjusted to take into account the changes in the national distribution of the population of Internet users (particularly neces-

[1] This project received support from the Ministry of Research through the National Telecommunications Research Network (RNRT).

sary in a high-growth market). The quality of information about Internet usage was guaranteed by collecting all the browsing information from the user workstation. Indeed, the diversity of the sites and activities available on the network meant that it was not possible to rely on users' declarations to describe their practices. Internet audience measurement panels were therefore an essential source of representative and exhaustive knowledge on Internet practices.

When our research project started in 1999, the NetValue panel seemed to us to be the most complete source. Indeed, the other audience measurement companies only monitored visits to Web sites and took no account of other uses such as e-mail chat, or file downloading, which nevertheless represent a growing proportion of Internet usage in terms of both volume and time spent. NetMeter, NetValue's audience measurement system, records all protocols other than *http*: e-mail, peer-to-peer software, instant messaging, network games, chat and streaming. When studying usage from a sociological point of view, the relationship between interpersonal communication and information browsing via the Web seemed to us to be central.

One initial methodological choice involved working with cohorts. Audience panels were used to produce monthly statistics for Web sites audience but were not designed to study long term trends. In the context of our partnership with NetValue, and subsequently in the SensNet project, our approach aimed on the contrary to study longitudinal transformations of practices within a closed population. To do this, we worked with cohorts, i.e. a set of households and individuals monitored over a long period. There is of course a limit to use of the cohort approach in a fast-growing market because a cohort becomes less representative as the months go by. This is why we chose to redefine a new cohort each year. We therefore simulated the possible trends in Internet practices once the market has reached saturation point. Some of our results seemed counter-intuitive due to the traditional confusion between a cross-section and a longitudinal analysis. We thus showed that the use of search engines was gradually falling, whereas the providers of engines were observing usage of their tools increase greatly (due to the increase in the number of new Internet users).

The second methodological choice involved working on qualified panels. Unlike site-centric approaches which have great difficulty in identifying the people behind IP addresses or cookies and find it even harder to categorize them, audience panels provide access to information about all the individuals in the household, as any connection requires authentication from a list of users (all members of the household panel). Individuals are described by the conventional socio-demographic variables concerning them and their household: age, sex, occupation, socio-professional category of the head of household, composition of the household, place of residence, income level, etc. Households were also described by their audiovisual and communication equipment ownership. Finally, information on their Internet connection was collected such as date of first connection, type of Internet connection (PSTN, DSL…), and place of connection.

Using the NetValue France panel, we defined a cohort composed of Internet users present in the last two months of 1999 and active at least once in 2000. This cohort was monitored throughout the year 2000 and the same approach was adopted to set up a cohort in 2001 and in 2002. A sub-cohort of about 600 individuals has also been monitored for almost three years (January 2000 – October 2002).

II.2. Possible strategies for collecting usage data

Several strategies have been determined for collecting usage data in “natural conditions”. They correspond to different positions of the probe in the technical Internet or Web activity processing chain.

The first approach is external to the Internet infrastructure and involves using video recordings of the user and screen. This is the method chosen for example in [7]: for 10 days, the participants in the experiment were invited to start a camera when using the Internet as well as to make comments on their actions to facilitate the interpretation of the recording. This system does not specifically collect traffic data but enables user behaviour interacting with the GUI[2] to be collected. This method provides very detailed data, although it is difficult to implement and does not allow massive representative long-term studies to be carried out on large panels of Internet users.

To collect actual traffic data, i.e. time-stamped records of user actions and network events, software components are needed. This solution usually involves positioning a probe in the processing chain going from a client workstation to remote servers.

In the first case, a probe may be integrated into a client program itself, for example, a Web browser. The first user-centric Web usage study [8] implemented this approach by modifying the XMosaic browser, as did [10] and [18]. These solutions allow very detailed recording of user activity in interaction with the GUI. The data collected loses in coverage what it gains in accuracy: only the software for which the component was developed is tracked, but the level of tracking is very detailed. A lighter version of this strategy centred on the client software used involves using existing data recording functions. For example, chat clients usually have built-in log capabilities. As for web browsers, the History function might be used to track user activity on the Web [9].

At the other end of the chain, network metrology tools enable positioning at intermediate points between the user workstation and servers: proxy servers, routers, etc. The client workstations are identified by their IP address. Non-intrusive probes examine IP packet headers and can measure the volume exchanged by protocol (by port number, more precisely). It is therefore possible to obtain precise, time-stamped information on the types of protocol used and volumes involved: Web, messaging, peer-to-peer, etc.

Finally, certain systems fall between these two types of solutions. They involve positioning the probe on the user’s workstation: all communications between the machine and the network can be tracked and specific users can be identified, rather than just the usage on a workstation in general. The probe must then implement specific software modules for each protocol to identify, analyse and extract the relevant information: for the outgoing mail service for example (SMTP protocol), the probeidentifies and extracts data such as message recipients, type of attached files, etc. Of course, this is easier for standard protocols, as for most of those used on the Internet (HTTP, POP, SMTP, NNTP); without this, analysis will prove more difficult and require retro-engineering to decipher client-server communication modes for proprietary protocols.

[2] Graphical User Interface.

II.3. The NetMeter User-centric Probe

As part of the aforementioned SensNet project, we used data collected by NetMeter, a technology developed by NetValue. Internet activity is monitored in real time using a software probe installed on each panellist's computer.

The information is analysed by identifying the different users in the household. NetMeter is unobtrusive proprietary panellist software which "silently" gathers and relays Internet usage data (such as use of chat, e-mail, instant messaging applications, forums, audio, video download and peer-to-peer applications) which passes through the TCP/IP layer. NetMeter is a small computer application, and has no impact on regular use of the panellist's computer; it automatically starts up when the computer is booted up. Data is collected and analysed at the network layer level, which stands between the different applications accessing the network and the remote servers. Regularly, data recorded by NetMeter is sent automatically via an Internet connection to a server without disturbing the user. This data is then validated and loaded into a database.

NetMeter is a user-centric probe: when recruiting the panel, each individual of a household is listed. This list is a parameter of the probe: it is submitted in a popup window on system startup and after more than 30 minutes of inactivity on the computer. This mechanism allows user identification and guarantees good reliability of individual data. We can then determine the network applications used and, for some of them (Web browsers, e-mail clients, etc), network flows are analysed more precisely so that information like URLs requested on the Internet, the recipients of an e-mail or the title of a newsgroup, is identified and sent to the collection servers.

The collected data contains, for each protocol, tracking data on each query, i.e. the exact time of the action (to the nearest second) and information specific to the protocol used, for each user. In addition, the names of executables accessing the network through TCP/IP are collected: iexplore.exe for Internet Explorer, msimn.exe for Outlook Express, etc. This enables user applications to be detected, and provides elementary data for protocols which are not analysed in detail. For example, the executable cs.exe corresponds to playing "Counter Strike" on a network. Even without analysing in detail the content of exchanges in the protocol used by the game, one can determine if a user is playing, when and for how long.

In terms of the protocols analysed, each of them sends information specific to it. In the case of Web traffic (HTTP protocol), not all queries made by the browser are recorded: image-type files are filtered out (GIF, JPEG, PNG formats, etc) when they are part of a page as is the case for the immense majority of Web pages. For Web traffic, the probe collects the following information: URL requested, referer (URL of the page containing the query or empty in other cases), HTTP return code and size.

For conventional e-mailing, the NetMeter probe analyses POP (receiving) and SMTP (sending) protocols and extracts the following information from each message the anonymous addresses of all correspondents listed in the message as well as their status, the subject and size of the message, the date of receipt on the server, the list of attachments and their type, although not the content for obvious reasons of confidentiality.

Finally, the probe does not analyse the content of data in other protocols but records raw information about exchanges with the network by applications with access to it. For each socket opened by an application, the probe records the name of the application, the time when the socket is opened, the duration of opening and the volumes of data uploaded and downloaded.

Using the NetValue panel data, which we have been able to process over a long period of time, we have a representative population of Internet users with detailed information on each of them and a fine-grained description of Internet practices.

III. FROM TRAFFIC DATA TO USAGE DESCRIPTORS

The Internet allows very different types of activities. These activities may use specific protocols (such as POP/SMTP for e-mail.), they can also use the http protocol, and the service is then accessible to the user via a Web browser (e.g. WebMail). In order to obtain data usable for sociological analysis, some processing of the collected raw traffic data is needed. This implies two things. Firstly, that traffic data should be processed to attach information on services and application fields. Secondly, this means that the discontinuity generated by the technical point of view of the data must be overcome to maintain the ergonomic continuity experienced by the user. For example, e-mail messaging is possible both through an e-mail client – using POP/SMTP protocols – and a Web interface.

One of the main problems we had to cope with was the categorisation of Web pages visited by our panel. For this purpose, we developed a software called *CatService* to categorise Web pages in terms of type of site they belong to (generalist portal, e-commerce site…), services they propose and topics they deal with.

III.1. Traffic Data Enrichment

III.1.1. The CatService platform

The Web data available from the NetMeter probe is provided as flat lists of URLS. To describe the usages and services which Internet users have accessed, content information needs to be attached to these URLS (theme, function, etc). The volumes handled mean that manually examining each page is out of the question. The 3,400 users analysed in the 10 months of 2002 visited almost 6.7 million different pages. We therefore implemented two strategies to automate this task. Firstly, particular types of sites and pages were categorised using specific rules and, secondly, Web directories were used as a resource for global classification of Web objects.

The CatService platform provides qualification of the visited URLS in terms of types of site and service. There are five levels of qualification:

- Site type: defines the type of site or content accessible on the site, for example "generalist portal", "WebMail" site, "e-commerce", etc.
- Portal: the site or portal to which the URL belongs. The system enables portals distributed over several domain names to be grouped under one single entity.

- Supplier: the provider of the service which may be called by the portal: for example, during the period of our observation, Google used to be the provider of the search engine service of the Yahoo portal.
- Service: within a given site type, a matrix of services proposed is defined and applied to all sites concerned. This means that comparable categories can be created within an analysis of a particular type of site and you can go beyond the headings defined for each site.
- Sub-service: the service may be divided into sub-categories. For example: the “search engine” service covers Web pages, images and contributions to discussion groups, as well as access to an advanced search page.

To operate, *CatService* requires a set of pattern matching rules, built with the help of a formalism based on regular expressions. These rules enable us to associate a class of URLS with a given portal-supplier couple, a service and a sub-service. These rules are constructed manually after examining the different URLS in a portal and verifying the page content to which they point. Regular expressions distinctly carry the domain name and the rest of the address and may be enriched with a “negative” regular expression which excludes the URL fulfilling it from the result. In addition, two specific processes are operated on certain types of service:

- For URLS that correspond to queries in search engines, a procedure extracts and standardises keywords in the query and also identifies browsing in the subsequent results pages.
- For URLS that access WebMail services, the tool identifies, where possible, login, message reading and writing actions.

An example of search engine rule will illustrate this mechanism:

If	RegExpHost	^(www\.)?google\.(com|fr|be|ch|de)$
	RegExpReste	(search|custo|advanced_search)
	KeyWord	(&|\?)q=
	Browsing	start=
Then	Portal	Google
	Supplier	Google
	Service	Search Engine
	Sub-service	Web

The reference base and rules are built manually by the users of the *CatService* application. This work is time-consuming, but *CatService* will then enable the share of each service used to be determined with precision, going beyond global audience measurement at portal or site level.

Categorizing services is of great value to our work, in particular for generalist portals. These sites represent the majority of Internet audiences, and *CatService*’s detailed description not only distinguishes, in each portal’s audience, between the different services used (search engine, WebMail, etc.), but also makes these elements comparable between different portals, as we have done on generalist portals in 2000 [3]. It also introduces an important services concept and enables a distinction to be made between pages whose textual content takes precedence and those where the function (the proposed service) is more important from a descriptive point of view. For example, it seems more relevant for analysing Yahoo pages, from a user point of view, to remember that one URL supplies continuous information rather than examining the content of information supplied on the page, which is constantly changing.

From this point of view, although *CatService* does not describe the entire Web, the choice of manually-categorised sites fulfils the need to describe those most visited by panellists. With the "generalist portal", "search engine", "WebMail", "homepages hosting" and "e-commerce" categories, it covers the most popular Web sites. The description of browsing in terms of services is therefore given a broad, solid base which corresponds to the main Internet activities: information, communication, e-commerce, banking services and leisure. In terms of coverage, applying *CatService* to the base of the SensNet project 2002 panel, the categories obtained describe about 30% of the URLs seen by the panellists. It is above all in terms of sessions that it shows its usefulness: *CatService* describes the pages visited in about 80% of sessions.

In addition, *CatService* allows the study of specific services at a more detailed level. Extracting keywords from queries sent to search engines enables advanced studies of the usage of different search engines and the formulation of queries to be carried out[3]. Semi-automatic categorisation appears to be a very productive and effective approach.

III.1.2. Web directories as a resource for Web pages categorisation

In addition to qualifying Web pages in terms of services, we need a method to qualify visited contents from a thematic point of view. The approach used to describe the content behind URLS involves using external data. We used the description of pages or sites in Web directories which may be compared to structured textual meta-data.

As one of the tools for searching content and services on the Web, a directory provides the user with a hierarchical classification of sites grouped under thematic categories. Unlike search engines, Web directories give a universal description of reference sites, and provide users with a commented classification of them, organising them into categories and sub-categories.

The aim here is to use the textual description of the site or the page indexed in the directory and its position in the categories and sub-categories to characterise its content thematically and functionally. This method of content characterisation has several advantages: first, a Web directory proposes a kind of categorization of the "Internet World", which is far from being perfect, but these categories can be used for classifying and labelling Internet Uses as we observe them in our panels. Besides, site and page descriptions are verified manually by the directory indexers: they should therefore be fair and accurate.

We have carried out an extensive comparative study of 8 French-speaking Web directories, including seven generalist directories and one specifically dedicated to homepages [2]. During this study, we developed a software to collect, for each studied directory, information on its structure (category and redirection structure) and the sites that it indexes (URL, title, description). The 8 directories studied contain almost 421,000 single sites or pages indexed. When we project these directories on the Web pages seen by our panel, we see that individual coverage rates of directories in 2002 were quite similar for the seven generalist directories,

[3] See the study carried out in [1], particularly showing the specificities of search engines depending on the queries sent to them and the profile of the users.

varying between 26% and 32%. We can therefore suppose that directories mainly index popular sites which concentrate large amounts of traffic. If we consider now all the URLs described by directories, the overall coverage of browsing is relatively good: 48.3% of the 6.7 million unique URLs visited by the 2002 SensNet panel feature in the eight directories, representing 42.5% of the 27.2 million pages visited. This broadly satisfactory coverage of the pages visited by the directories allows us to use them to describe and characterise Internet user browsing. We thus chose to analyse Web sessions having more than 50% coverage by the Web directories. For each page described in the directory, we noted the corresponding top-level categories; we then looked in the session at the time spent on each category and chose the most representative in the session.

Finally, both Web directories and *CatService* are used for efficiently describing Web contents. With this combined use of *CatService* and Web directories, the rates of session coverage by the descriptions are greatly improved. Overall, 48% of the observed traffic is described in terms of duration, and 53% of the sessions are described for more than half of their duration. Amongst these well-covered sessions, between 30 and 35% of them are usually described completely (about 15% of all sessions). Such coverage rates thus allow us to perform qualitative analysis of Web uses.

III.2. Defining Internet sessions

The data enrichment process described above allows us now to define multi-protocol Internet sessions. Observation of Internet users and qualitative and ethnographic studies carried out on Internet usage have shown considerable interpenetration between activities on the Internet: Web browsing, e-mail, forums, chat, instant messaging (see [4]). We have built an Internet session that includes all types of activity on the network. This session concept is central for us. Indeed, it is the unit of measurement chosen to compare types of activity that leave very heterogeneous traces on the network.

The technical separation into protocols makes little sense from the user's point of view. They coexist on the computer and interact in terms of the content proposed (a site proposes contact via an e-mail address, a post to a newsgroup contains a link to the contributor's homepage, peer-to-peer software uses web pages for updates, etc.). This significant interpenetration of the different services available on the Internet implies that data collection systems must adopt a comprehensive approach which is almost naturally in opposition to the technical imperatives that they face in analysing protocols. Behind the attractive user interfaces and the increasing interoperability of Internet tools and services, we discover a world of technical systems whose interactions are complex. A traffic data collection system must integrate this complexity and discontinuity whilst preserving the continuity of the system from the user's point of view.

To do this, we first qualified the traffic observed in data on protocols other than the Web and e-mail. We identified different application systems communicating with the network: instant messaging, games, peer-to-peer, etc. We then manually linked the executables listed in the data to specific, identified software and classified this software into an application category system. When observing the results of this categorization, we can see extensive interweaving of the different tools shown in the example presented in Figure 1 below. Adopting a user point of view means using all this activity to identify sessions. So, within the SensNet project, a specific methodology has been set up which includes all protocols used to identify

Internet sessions. This modifies, often significantly, the measured duration of sessions as well as their number and has an influence on service usage measurements during a session.

Date	Application	Duration (sec.)
01/01/2002 09:59:23	Mail – reception	0
01/01/2002 09:59:23	Mail – reception	0
01/01/2002 09:59:23	Mail – reception	0
01/01/2002 09:59:23	Mail – reception	0
01/01/2002 10:00:03	Web	513
01/01/2002 10:00:34	Instant messaging	1
01/01/2002 10:00:35	Instant messaging	5
01/01/2002 10:00:56	File transfer	4
01/01/2002 10:03:09	Mail – reception	0
01/01/2002 10:08:14	Mail – reception	0
01/01/2002 10:08:36	Web	172
01/01/2002 10:11:28	Web	600
01/01/2002 10:28:36	Mail – reception	0
01/01/2002 10:31:41	Instant messaging	1
01/01/2002 10:31:42	Instant messaging	8
01/01/2002 10:31:59	Mail – Reception	0
01/01/2002 10:32:41	Web	79
01/01/2002 10:34:57	Web	0

FIG 1. – Example of a multi-protocol Internet session

Légende française

The main technical problem was to find a heuristics for defining sessions boundaries, since in our data, there is no explicit indication of when Internet usage starts or finishes: we see traces of activity interrupted by varying periods of inactivity. The issue is to define which inactivity duration will determine whether a session is finished, so that subsequent tracking data belongs to another session.

In our data, the average time between two consecutive events (web page consulted, e-mail received or sent, etc.), being 12 minutes, three hypotheses have been tested: declaring the session end after 15, 30 and 45 minutes of inactivity. As the parameter differences tested (average session duration, number of sessions per month, etc.) stabilised between the 30 and 45 minute hypotheses, the limit finally chosen was 30 minutes of inactivity before declaring a new session.

IV. RESULTS: USAGE PROFILE AND TRAJECTORIES

In this last section, we present some of the results produced using the platform described previously. We will start by presenting a typology of Internet users based on Internet activity (Web, mail, search engine, other protocols); we will then present a typology which takes into account the content of the visited sites.

The results today have an historical value. They provide a map of Internet usage at a time when broadband was not widespread and when usage involved above all text documents which were at best illustrated. They show the changes in use between 2000 and 2002 at a period of transition for the Internet in France, where access to broadband has expanded greatly. The development of the production, exchange and consumption of musical and video documents has transformed the range of uses.

These results also have a methodological value: they illustrate the added value of longitudinal work on cohorts of Internet users representative of the population connected. This methodology has proved essential for observing the effects of learning and appropriation of the Internet. It also helps to detect weak signals and the profiles of minor users in terms of usage, who are usually hidden in the global audience measurements by more intensive users.

IV.1. Typology and changes in online usages

IV.1.1. Two very different profiles in 2000

We have built up a typology of Internet users in the year 2000, recording information on types of usage (Web, mail, forums, chat, instant messaging, search engines, etc.). Using a matrix describing sessions on the basis of a certain number of criteria, particularly inter-personal communication practices, the following approach is used:

1. First of all, a *session profile* is created for each session according to the presence or absence of different types of activity in that session: communication, Web, instant messaging, games, peer-to-peer, etc.
2. Each internet user is described by her/his session profiles as well as global variables concerning her/his intensity of use across the different application fields.
3. Internet users are classified according to these profiles, using principal component analysis and an ascending hierarchical classification

Among the one thousand or so Internet users observed, we found two major usage profiles: on the one hand, light and very light users of the Internet, who represent 47% of the cohort but only generate 15% of sessions throughout the year and, on the other hand, the heavier users (53% of the cohort and 85% of the sessions registered). This contrast reflects very different Internet learning curves. The results show that Internet usage declines over a period of time for the first group whereas it increases for the others. Thus, two Internet usage trajectories can be seen, with a decline in usage among the lighter users and growing usage among the others. Light Internet users are characterised by their lower use of e-mail. [13] have shown that use of e-mail and therefore access to a network of Internet correspondents was the main factor in stabilising Internet practices and transforming them into a "routine". We see here that the absence of correspondents is accompanied by low levels of usage and is often a sign that the person is likely to stop using the Internet altogether.

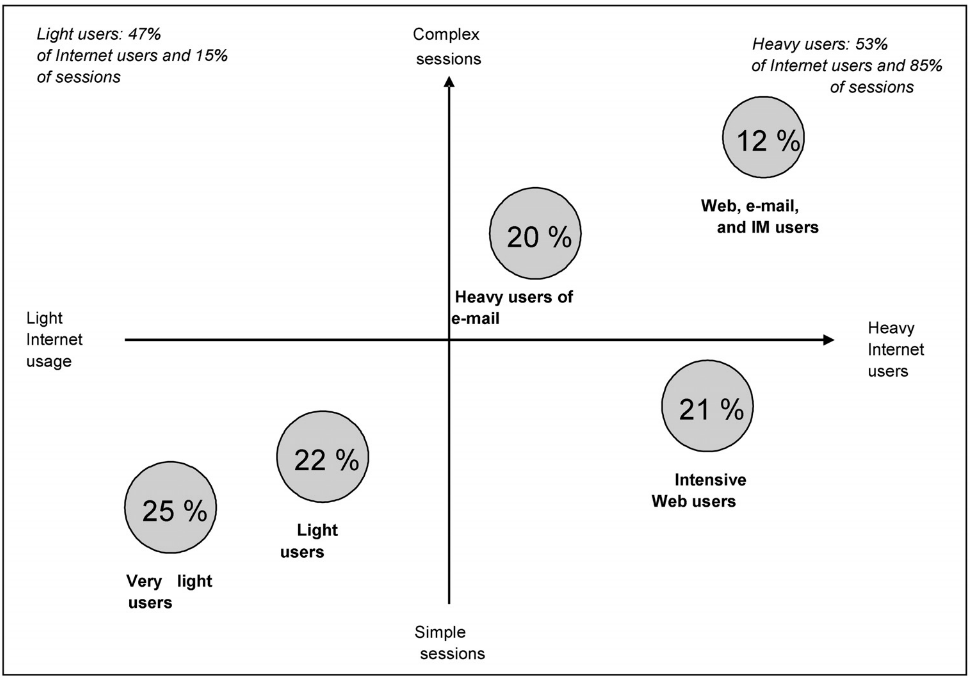


FIG 2. – Typology of Internet users in 2000 based on the profile of their Internet sessions

Légende française

Let us now focus on the heaviest users whose practices, unlike the others, have reached a form of stabilisation. There are three groups. The first, representing 22% of the cohort, includes intensive Web users for whom interpersonal communications, and therefore writing, are secondary[4]. For the other two groups of users, inter-personal exchanges take precedence over Web browsing. The second group, representing 20% of Internet users, comprises users of e-mail[5]. Although e-mail is a core part of their usage, they nevertheless use the Web,

[4] They browse the Web in almost all their sessions (85%), only access e-mail in 30% of their sessions and never use instant messaging or chat services.

[5] They look at their messages in almost two thirds of their sessions and Internet browsing only involves half of them. Live conversation is not part of their usage.

although less intensely than the first group. The third and last group (11% of users) includes individuals who have conversations on the network and therefore use communication tools at the same time as other users (chat and instant messaging).

We are confronted with a tiered model where new usage does not lead to the abandonment of the others. The first group essentially browses the Web and gives little importance to discussion. The second group, while it does not ignore the Web, gives greater importance to e-mail combining browsing and correspondence. The third group combines browsing and e-mail like the previous group and adds live conversations. The way in which they take advantage of the potential uses of the network varies considerably from one group to the next.

Users in the second group, showing a high level of e-mail usage, tend to come from a higher social class, while users of electronic conversation tools (*chat* and instant messaging) come from more modest backgrounds. A certain number of hypotheses can be advanced such as the valorisation of writing and books in more affluent environments leads to a disregard for these types of communication with no “memory”, which use a form of writing very different from accepted usage; the poverty of the content and absence of “informational” finality also makes these exchanges suspicious. Conversely, this writing without “memory”, with its framework of local standards separate from the dominant norm, enables lower social groups to overcome the barrier of writing.

Another notable contrast is between younger and older users. Communication tools play a central role in the Internet practices of younger people, but their real specificity lies in their ability to articulate and combine very diverse uses. This dexterity in front of the screen, chaining a variety of tasks, seems to be what distinguishes them from older users.

Finally, women are almost absent from the group of intensive Web users although they are present in the other groups where inter-personal exchanges are important. On the Internet, as elsewhere, women’s strong commitment to maintaining their relational network is clearly visible.

IV.1.2. Changes in usage in 2002

Between 2000 and 2002, the panorama of Internet users in France changed considerably. Firstly, subscription levels rose sharply, and from the offer side, a considerable number of new online services appeared (e-commerce, e-banking, peer-to-peer…). Secondly, this development was supported by the deployment of high-speed lines, particularly using ADSL technology, which took off at this time. In this context, usage showed significant changes, accompanied by a number of constants.

The greatest difference between 2000 and 2002 was the widespread distribution of peer-to-peer software. Peer-to-peer software was only used in 2% of sessions in 2000 (Napster was almost the only software used) but was present in 6% of sessions in 2002 (with Kazaa well ahead). At the same time, traditional messaging and the Web continued to attract the majority of users and to constitute the base of Internet usage. To show the complexity of these changes, we built a new typology to describe usage in 2002 based on the intensity of usage. Each Internet user is described by 10 discrete variables relative to their usage of each type of protocol on the Internet. Multiple component analysis on this basis enables us to identify four distinct major classes of users, shown in Figure 3.

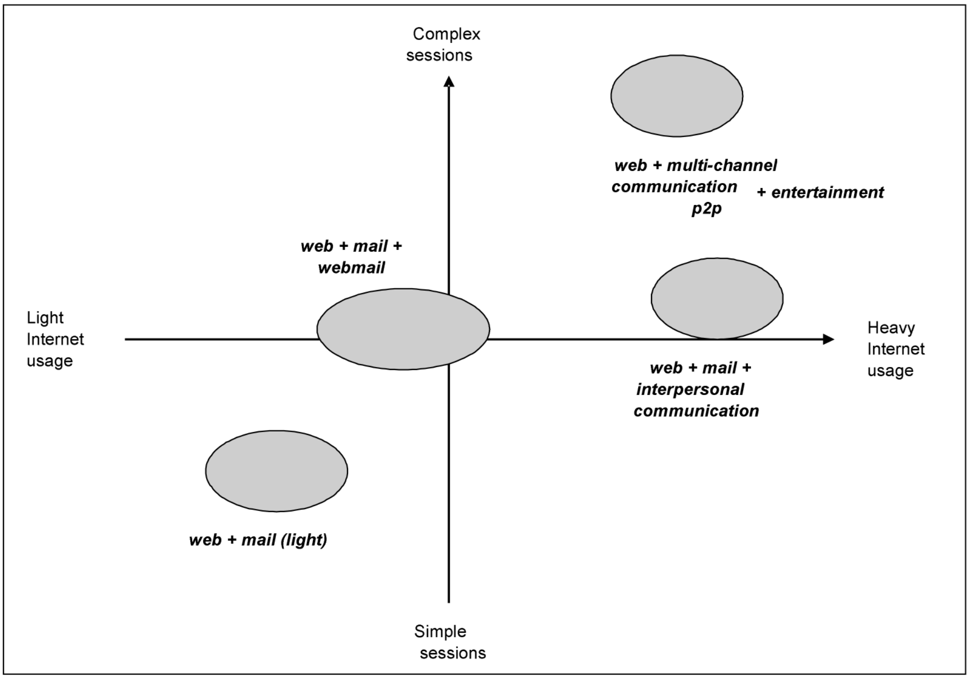


FIG 3. – Typology of Internet users in 2002 based on the profile of their Internet sessions.

Légende française

A first group of users is distinguished by the low intensity and lack of variety of its usage. Among these occasional users, who represent 33% of the cohort, only 2.9 sessions per month were detected on average (median: 1 session), compared with 145 for the entire cohort (median: 70). The services used are limited to the “fundamentals”: Web and e-mail.

The second group represents 35% of the cohort and has uses which are both more intensive and more diversified. This class of “ordinary Internet users” accounts for 9.3 sessions per month on average (median: 7.2) and has medium Web, e-mail and WebMail activity; a little downloading and audio/video content supplement this picture which contains no peer-to-peer or advanced communication tools (chat, Web Chat, forum).

These services are the prerogative of the third group of users (22% of the cohort) whose use of interpersonal communication tools is particularly intensive. At the same time, Web usage is also extremely high among these “communicating Internet users”, who use messaging and chat as much on the Web as with their specific tools.

The last group, 10% of the cohort, is far more oriented towards entertainment activities: peer-to-peer usage, absent or very light in the other three groups, is particularly intensive here, along with audio/video services and file downloading. In terms of instant messaging, these heavy Internet users are distinguished not only by a high level of usage but also by a shift from WebChat to chatting via a specific program (ICQ, Messenger, etc). Although it is not a determining factor, network gaming is also very widespread in this group where it concerns 40% of individuals, compared to 10% in the cohort as a whole. Finally, these intensive users are those who use the Web the most, with 39 sessions per month on average for each panellist.

In 2002, the 600 users in the cohort monitored for three years were a sub-group of the SensNet 2002 cohort. It is interesting to note the group in which these older users ended up in 2002: it might be supposed that more than three years of Internet usage would imply anchoring and intensification of their practices. It seems however that this is not always the case: the distribution of Internet users in the 2000-2002 cohorts across the four standard categories generally follows that of the general cohort monitored in 2002 (see Table I).

TABLE I. – Distribution of Internet users in the groups

	Cohort 2002	Cohort 2000-2002
Occasional users	33.3%	24.2%
Ordinary users	34.8%	34.8%
Interpersonal communication	21.5%	27.7%
Entertainment use	10.4%	13.3%
Total	*100%*	*100%*

The “older” Internet users are more active than the others: they generally have a stronger presence in the “ordinary users”, “interpersonal communication” and “entertainment usage” groups than the 2002 cohort as a whole, to the detriment of occasional usage. However, a quarter of them are part of this last group, compared to the third in the 2002 cohort as a whole. This result shows that length of use is not a systematic driver of usage intensity and that using the Internet may be anchored in individual practices whilst remaining a resource used only occasionally.

IV.2. Learning the Web: personal territories on the Web

The traffic data shows us the sites visited by each panellist throughout the period and the context in which each of them appears. A map of the resources used on the Web can therefore be drawn up for each Internet user and the extent and the structure of this personal territory can be observed in the long term. For this, the 2000-2002 cohort gives us the essential depth of view to observe and analyse the composition and dynamics of these territories. We will use this data to examine this issue more closely.

IV.2.1. Variable intensities, constant habits

The 600 Internet users in the 2000-2002 cohort visited 192,000 different sites and portals during the 34 months of observation. However, not all these sites were seen by the entire cohort – in fact, far from it (see Table II): two thirds of the sites were visited by only one user, whereas another quarter were visited by between 2 and 5 different panellists.

TABLE II. – Cohort 2000-2002, number of different panellists per site/portal

	Number of sites	Share of the sites
1 visitor	127,357	66.3%
2 to 5	49,446	25.8%
6 to 99	14,818	7.7%
More than 100	378	0.2%

Only about 30 sites were seen by more than half of the cohort: this list includes the leading generalist portals and service providers (MSN, Wanadoo, Yahoo, Club-Internet, etc.) which have established themselves as crossroads and points of passage for going “elsewhere” on the Web (via search engines in particular) and propose additional communication services attracting a large number of users and winning their loyalty. The Website audience thus appears extremely fragmented: with the exception of the few points of passage common to the majority of users, everyone seems to be searching the Internet for content which interests them specifically. This is not really a surprise: when observing only 597 people, even for 34 months, with the varied nature of the cohort in terms of centres of interest, there is little chance to observe individuals visiting the same type of content.

A the single-user scale, the number of different sites and portals visited by each individual in the 2000-2002 cohort during the 34 months of observation varied greatly from 11 for the lightest user to 8,900 for the heaviest, who visited an average of 943 sites (median: 569). However, a user does not view all their sites equally. An examination of how often they are visited in different sessions by each panellist shows that, on average, three quarters of the sites visited by the cohort during the 34 months of observation appear in one session only. This dispersion is not linked to the intensity of use by each panellist. Intensive Internet users do not tend to see more new sites or have more frequent “browsing” practices than more moderate users limited to a few well-defined sites.

IV.2.2. A constantly-expanding corpus

On this basis, it might be thought that everybody has well-determined territories on the Internet and that, once the phase of discovering interesting sites has passed, the “corpus” stabilises around a few of them. In reality, each panellist’s corpus of sites expanded constantly during the 34 months of observation. On average, when half the sessions observed over the period had been completed, only 57% of the corpus had been explored and, when 90% of the sessions observed hade been completed, only 91% of the corpus had been consulted.

This almost linear link between the number of sessions and the number of different sites visited shows that the user continued exploring new sites in both the first and last months of observation. The large proportion of sites visited in a single session only cannot therefore be explained by an initial period of Web discovery, but seems to be an integral part of daily browsing practices.

In the end, users remain loyal only to a handful of sites. This notion should not be confused with the regularity and presence of a site in a panellist's sessions. The session must be linked to a specific activity and consequently regularity may be very different from frequency. A panellist may systematically visit the same site for a specific task which may only be required occasionally. The online tax declaration is a typical example of this situation, as the user only visits the Ministry of Finances' site once per year but may do so every year at the same time.

To analyse loyalty to sites, we examined the "span" of a site, i.e. the number of days separating the first and last visits to a site seen several times by the panellist in the 34 months of observation. In one third of cases, this period exceeds 30 days and only exceeds one year in one quarter of observations. An average of 72 sites per panellist have a span exceeding one year, with fewer than 16 sites for less active users and more than 100 for the most intensive. The most remarkable element here is that, despite the diversity and intensity of use, the number of sites seen at more than one year's interval is always proportional to the total number of sites visited by the user. For more than 90% of users in the 2000-2002 cohort, these sites represented between 25 and 35% of all those visited several times during the period and 7.5% of all sites visited

In fact, sites seen frequently and over a long period are rare. For those seen more than once in the three years by a given user, there is an average of 83 days between two visits (median: 34 days) and, for a quarter of sites, this interval is less than one week. Frequency and loyalty should not be confused and territory on the Web appears to be divided into three zones:

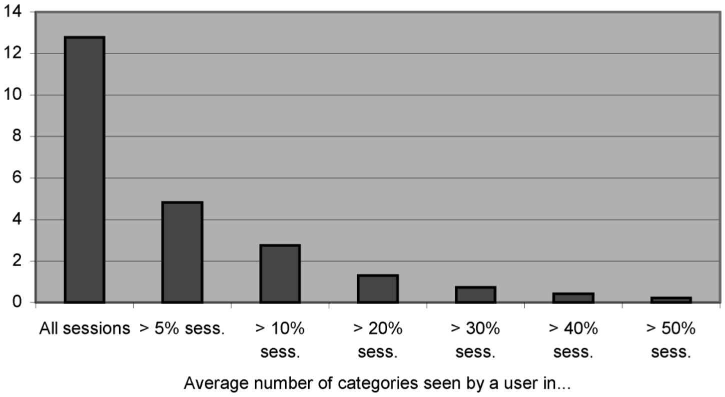
1. Sites only seen in one session, accounting for three quarters of the sites seen by an Internet user on average
2. Sites seen more than once but which do not win the user's loyalty These sites are active over short and medium-length periods, with an average of 80 days (median: 34 days), and appear in only two sessions in half the cases. On average, they account for 17% of a user's corpus and appear to correspond to passing activities for an individual or to "rejection after testing"
3. Familiar sites: seen over more than one year, they are visited preferentially by the user in a given context. This category obviously includes generalist portals and search engines, as well as services which win user loyalty, such as WebMail. It also contains specialised sites on certain subjects in which the user has a special interest. They do not systematically appear in many sessions – less than 5 in 50% of cases, but seem to be the user's preferred sites for a given domain or service.

IV.2.3. Thematic territories

The concept of thematic territories and preferential services on the Web echoes the definition of territories and usual places in terms of visited sites. Using descriptions of content which can be attached to the pages visited, we can map the themes and services accessed by each user and examine the frequency and regularity of access to this content in the sessions.

To do this, we describe each session by its CatService category or the first-level directory category in which the user spends most time in the session. As we have seen, in most cases this majority category covers most of the duration of the sessions as they are single-theme or single-functional sessions. We can therefore build a profile of the content visited by each user, corresponding to the share of the session spent on one content describer or another.

An examination of the profiles shows that users each have well determined centres of interest on the Internet which occupy most of their time on it. A strong concentration effect can be observed: firstly, although the number of descriptive categories ranges from 31 to 34 depending on the directory considered, only 12.8 categories on average describe all the sessions of each panellist; secondly, among this dozen categories to which the panellist devoted at least one session, two thirds were visited very occasionally and featured in fewer than 5% of sessions, whereas a single category describes more than 35% of the panellist's sessions on average (see Figure 4).



Key: for each user in the cohort, 13 different categories describe all their sessions on average, but only 5 categories are represented in more than 5% of sessions.

FIG 4. – Number of categories describing the panellist's sessions.

Légende française

Key: for each user in the cohort, 13 different categories describe all their sessions on average, but only 5 categories are represented in more than 5% of sessions

The number of different categories in which the panellist is interested is quite logically a function of that panellist's intensity of use. The more sessions there are, the more the panellist is likely to visit sites with varied content. So the least active quarter of the cohort saw a total of 5 different types of content on average over the 34 months of observation, whereas the most intensive quarter saw more than 18. However, it is noteworthy that the number of categories representative of more than 5% of sessions bares no relation to the intensity of use and remains around 5 for the cohort as a whole.

These figures confirm the idea of users creating their own personal space on the Internet. The observation regarding the visited sites also hold true for content. Although an individual may visit a wide variety of sites, the majority of practices remain centred on a limited number of services and topics. This fact encourages us to put into perspective, or at least specify, the concept of "surfing" as it is generally used. Although the Internet can give access to all types of services and all types of documents or topics, in practice, this possibility is only used on a macroscopic level as everybody creates niches around their centres of interest.

When studying personal web sites in 2000-2002, C. Licoppe and V. Beaudouin noticed that sites authors made efforts to offer their visitors communication capabilities (e-mail feedback, contact forms, forums, guest book, etc.) and were highly involved in interacting with their audience in order to improve the quality of their sites [14]. Today's blogs, wikis and social media provide users with adapted technical solutions to these needs. The evolution of the Web since 2002, when this behaviour was observed, has confirmed the importance of personal Web spaces in daily usage. After the decrease of portals conceived as a collection of wide-range information and services, the "portal strategy" is rising again with the help of highly customizable interfaces, with web sites such as NetVibes[6]. The two key technological points here are RSS/ATOM feeds for content retrieving, and Ajax technologies for their aggregation into an interface that fits exactly the user needs. In these Web 2.0 interfaces, the concept of "page", major in the first era of the Web, might collapse: browsers are moving towards interfaces for wide-range web-based application (office apps, Instant Messaging,…) interweaved in this interface. In this perspective, Web territories might be in the future focused on a central custom page which would be a kind of "desktop in the desktop", linked with varied services and sources of information, among which would figure the Web as we know it today.

V. PERSPECTIVES

The set of tools presented in this paper has enabled us to produce innovative results on Internet 1.0 usages, such as typologies of Internet practices in France, linking these types of practice to detailed profiles of users. We also combined content descriptions with topological and temporal indicators in a global approach to browsing and showed off the strong link between page content, browsing dynamics and users' personal territories on the Web [5]. Our platform was also used to carry out a large number of usage studies in the context of the SensNet project. The precision of the content descriptions as well as the upgradeability and scalability of the *CatService* module allowed us to perform fine-grained focuses on particular kinds of sites or services: use of Web search engines and of e-commerce sites [1] and [15], use of Digital Libraries [6].

In this article, we described the main modules of our platform for qualifying raw usage data collected on the user workstation. This type of processing is rarely described in detail in "Web Usage Mining" literature, even though it involves substantial technical difficulties and raises fundamental methodological questions. The choices made about the qualification of visited pages, or about defining concepts as fundamental as "site" or "session", have a substantial impact on analyses that may be carried out for Internet usage studies.

[6] http://www.netvibes.com

We have presented a system of usage measurement corresponding to the state of the Internet at the beginning of the decade, but it has now moved on. Indeed, mass availability of broadband access with contracts encouraging permanent connection, as well as the growth in downloading of audio and video contents, mean that the measurement tools will have to evolve. Although monitoring network activity on the PC allowed us at that stage of Internet development to capture a significant part of Internet uses, this approach is already less efficient given recent changes. The Internet service offering has expanded and outgrown the limited framework of the computer to provide bridges to other terminals such as mobile telephones, PDAs, digital music players, etc. The constant development of the Internet is forcing us to change the technical measurement tools at the same rate so that they can monitor usage as efficiently as possible.

In this context of extension of the Internet's "scope of action", the challenge is no longer to improve knowledge of Internet usage per se but to place these uses in the wider context of social and cultural practices. Online resources are now accessible from all types of devices and digital contents circulate on these different devices.

We are continuing our work in this perspective, setting up a representative panel for which we are collecting information on Internet and fixed/mobile telephone usage, as well as data collected from users on their social habits, their equipment ownership and their cultural consumption. We are working on technologies to measure not only network traffic but also overall computer activity, linking it to an active window and a specific application. This method means that overall user activity on the computer can be recorded and online activity can be analysed in the general context of digital activity.

With this tool, we are aiming to produce a study of Internet usage in all its complexity and in its longitudinal development, using panels representative of the population. As we mentioned in the introduction of this paper, the work carried out on Internet usage the work on telephone usage which also draws on traffic data. However, the Internet does not operate as a closed medium; the "virtual" Internet world is not separated from the "real" world as some early work would have had us believe. On the contrary, Internet usage is part of a much wider social context and cultural practices. Observation methods and tools must perpetually adapt to be able to report this dynamics.

Manuscrit reçu le
Accepté le

REFERENCES

[1] ASSADI (H.), BEAUDOUIN (V.), Comment utilise-t-on les moteurs de recherche sur Internet ? *Réseaux*, *20* (116). 2002, pp. 171-198.

[2] ASSADI (H.), BEAUVISAGE (T.), A comparative study of six French-language Web directories. in *ISKO 2002*, Granada, Spain, 2002, pp. 271-278.

[3] BEAUDOUIN (V.), ASSADI (H.), BEAUVISAGE (T.), LELONG (B.), LICOPPE (C.), ZIEMLICKI (C.), ARBUES (L.), LENDREVIE (J.), Parcours sur Internet: analyse des traces d'usage, France Télécom R&D, 2002.

[4] BEAUDOUIN (V.), VELKOVSKA (J.), Constitution d'un espace de communication sur internet (Forums, pages personnelles, courrier électronique.). *Réseaux*, *17* (97). 1999, pp. 121-177.

[5] BEAUVISAGE (T.), Sémantique des parcours des utilisateurs sur le Web, *Linguistique*, Paris X, Nanterre, France, 2004, p. 360.

[6] BEAUVISAGE (T.), ASSADI (H.), Uses of Online Digital Libraries – a User-Centric Approach. in *Eurescom Summit 2003*, Heidelberg, Germany, 2003.

[7] BYRNE (M.D.), JOHN (B.E.), JOYCE (E.), A day in the life of ten WWW users, *International Journal of Human-Computer Interaction*, 1999.

[8] CATLEDGE (L. D.), PITKOW (J. E.), Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, *27* (6), 1995, pp. 1065-1073.

[9] COCKBURN (A.), MCKENZIE (B.), What do Web users do? An empirical analysis of Web use, in *International Journal of Human-Computer Studies*, 2000.

[10] CUNHA (C.), BESTAVROS (A.), CROVELLA (M.E.), Characteristics of WWW Client-based Traces, Computer Science Department, Boston University, 1995.

[11] HAERING (H.), Les outils de mesure d'Internet, in *Les mesures de flux Internet*, Paris, 2003.

[12] LEBART (L.), La mesure des audiences en vue de la mesure des usages, in *Les mesures de flux Internet*, Paris, 2003.

[13] LELONG (B.), Thomas (F.), L'apprentissage de l'internaute: socialisation et autonomisation, in *CIUST'01, Colloque International sur les Usages et les Services des Télécommunications – e-Usages*, Paris, 2001, pp. 74-85.

[14] LICOPPE (C.), Beaudouin (V.) La construction électronique du social: les sites personnels. L'exemple de la musique. *Réseaux*, *20* (116), 2002, pp. 53-96.

[15] LICOPPE (C.), PHARABOD (A.-S.), ASSADI (H.), Contribution à une sociologie des échanges marchands sur Internet. *Réseaux*, *20* (116), 2002, pp. 97-140.

[16] O'REILLY (T.), What Is Web 2.0, O'Reilly Media, 2005.

[17] SMOREDA (Z.), LICOPPE (C.), La téléphonie résidentielle des foyers: réseaux de sociabilité et cycle de vie, in *2^e Colloque International sur les Usages et Services des Télécommunications*, Bordeaux, 1999.

[18] TAUSCHER (L.), GREENBERG (S.), How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human Computer Studies*, *47* (1), 1997, pp. 97-138.