FUNCTIONING AND USE OF AN INSTITUTIONAL REPOSITORY[1]

Frederic Merceur
Ifremer – Bibliothèque La Pérouse
BP 70, 29280 Plouzané, France
frederic.merceur@ifremer.fr

Abstract:
In August 2005, Ifremer launched its institutional repository: Archimer. Today, Archimer provides free access, on the Internet, to more than 5,000 documents, including more than 80 % of the publications co-written by Ifremer since the creation of the repository.

Following a reminder of the publication harvesting and recording methods, this document assesses the use of Archimer. It analyses, amongst others, the progression of the number of interrogations and the differences observed between the uses of the different types of works (publication, thesis, internal reports...).

This study also demonstrates the predominance of the search engine Google in the access to the documents and underlines its consequences in terms of Internet visibility.

Keywords:
Open Access, Institutional Repository, Archimer, Google, Statistics

[1] Most of this text has already been published in French and can be accessed at the following address: http://www.ifremer.fr/docelec/doc/2008/rapport-4632.pdf

Historical Reminder:
In 2003, the head of Ifremer expressed the wish to see its scientific publications be given more visibility, and especially, at first, its theses and internal reports.

As an answer, the library staff developed the "Thèses & rapports en ligne" (online theses & reports) Website which enabled the recording and publishing of electronic documents. This first Website was launched in May 2004.

At the beginning of 2005, the library staff offered the head of Ifremer to improve this system in order to enable the recording of three new document types: conference

proceedings, activity reports generated by Ifremer departments and/or laboratories, and, above all, publications written or co-written by members of the Ifremer staff.

In the framework of this diversification, we suggested that this initiative should be in line with the "Open Access" movement, and that the new version of the "Thèses & rapport en ligne" Website should be presented as Ifremer's institutional repository. In August 2005, this new version, renamed Archimer, was launched on the Internet. Archimer is accessible at the following address: http://www.ifremer.fr/docelec/.

Document Recording Conditions:
Documents are loaded in Archimer by the Institute's librarians who are in charge of:

- Entering metadata,
- Filing documents according to specific topics (ex: biology, aquaculture, fishing …),
- Full-text formatting and converting into PDF if necessary,
- Transferring full-texts to Archimer's server.

Publication Recording:
Some of the authors tell us which publications they would like to be published in Archimer. In that case, we check which rules have been set up by the publisher of the publication as far as self-archiving is concerned. If the publisher authorises self-archiving, we provide the author with the information we need to record those publications.

However, in order to record and broadcast a greater amount of publications, we do not only count on spontaneous submission by Ifremer authors, but we handle the following watch and collecting:

- Each week, we search for the publications written or co-written by Ifremer in the Web of Science®,
- Then for each publication, we consult the publisher's policy on the Sherpa/Romeo Website. If the publisher's policy is not specified on the Sherpa/Romeo website, or on its own website, we systematically contact the concerned publisher to ask permission to record the articles into Archimer,
- If the publisher allows the self-archiving of its own PDF files (ex: EDP Sciences, The Company of Biologists ...), we download ourselves the PDF file of the concerned article from the editor's website and then record it into Archimer,
- If the publisher restricts the right of self-archiving to the author's final manuscript of the publication, we contact the authors of the publication to request a version using some automated tools developed by Ifremer. If they can provide us with this version, we issue ourselves a PDF file from the files sent, and then register it into Archimer. In most cases, these publications are sent
-

- under multiple files (one file for the text and other files for tables and images). We merge all these files, we create a cover sheet with the references of the
- article, we use a minimalist layout of the text before turning it into PDF and optimizing it for a better visibility on the Web (ex: http://www.ifremer.fr/docelec/doc/2008/publication-4501.pdf)

As a consequence, we contact authors all through the week to collect their publication. A recovery system, also automatic, enables us to remind the authors that we expect their paper for Archimer. In most cases, we need to contact the authors 2 or 3 times (and sometimes more) before obtaining their publications. This harassment policy causes few calls from scientists fed up with these reminders, but fortunately, this work is generally very well received.

Obviously this method is not perfect: it is costly in time (1 working day for 10 publications), fragile because of the financial difficulty to hire staff and the collection of documents based only on spontaneous deposit does not work (conference proceedings, internal contract reports ...).

But this method enables an important collection of publications referenced in databases. It helps to overcome:

- The authors' lack of time,
- Their lack of immediate interest. If the promotion of free access is not necessary for physicists for whom ArXiv has become unavoidable, this is not the case for other fields like, for example, life sciences - in majority at Ifremer. For these scientists, who use tools like the Web of Science® or ScienceDirect®, Open Archives are not yet working tools,
- Some scientists' (this is particularly true for people working on life sciences, in majority, at Ifremer) lack of computer knowledge. A scientist answered our request for his final manuscript saying: "The version is not publishable (heavy file, figures and text separated)",
- Their ignorance of copyright policies

This method has also other advantages:

- Saved documents are optimized for better visibility on the Web,

- These personalized contacts with authors enable us to increase their awareness of the practical aspects of the Open Access movement faster and especially to make them understand the importance of their final manuscript,
- The recording of an important number of documents will enable a fast collection of a critical amount of documents. This amount, will mechanically enable a better visibility of the project, and could quickly give us:

    - The recognition of Ifremer's authors (more spontaneous deposition...)
    - The recognition of Ifremer's leadership (more resources...)

In terms of results, this method enables us to collect **nearly 80 % of the international publications co-written by Ifremer since the launch of Archimer.** Since August 2005, Ifremer has co-published 1,480 articles referenced in the Web of Science® database. Out of these 1,480 publications, 1,180 are now in free access in Archimer, that is to say about 80 %.

Thesis, Report and Proceeding Recording:
Concerning theses, conference proceedings and internal reports, we hoped, when the system was implemented, that authors would submit their works themselves. Unfortunately, even if the number of spontaneous submissions is increasing, documents collected through this method are still a minority, if not marginal.

In order to increase our thesis collection rate, we contacted the head of the human resources who now forwards us, on a regular basis, the list of the last PhD students who defended their dissertation proposal. This information allows us to contact them and offer the recording of their thesis into Archimer. **These personalized contacts enable us to collect about 90 % of the theses funded by Ifremer.**

In order to increase our report (internal reports, contract reports, monitoring report…) and proceeding collection rate, we started to analyze, in 2008, on a trial basis, the activity reports generated by the different Ifremer departments. These activity reports provide us with each department's annual publications list.
Concerning conference proceedings, this analysis helped us find out that most conference contributions today are limited to the creation of a slide show or a poster and do not involve the writing of a formal document. Consequently, a large majority of these documents do not fit in Archimer.

Regarding reports, a confidential visibility mode, available in the upcoming version of Archimer, will enable the collection of a greater number of them.

State of Affairs

Number of Recorded Documents:
On September 16, 2009, around 5,900 documents have been recorded in Archimer with 5,400 of them accessible freely on the Internet. Figure 1: Evolution of the number of documents available in Archimer
shows the increase of the number of documents recorded in Archimer since October 2004. The following tables present the distribution by document type and by subject of these 5,400 documents.

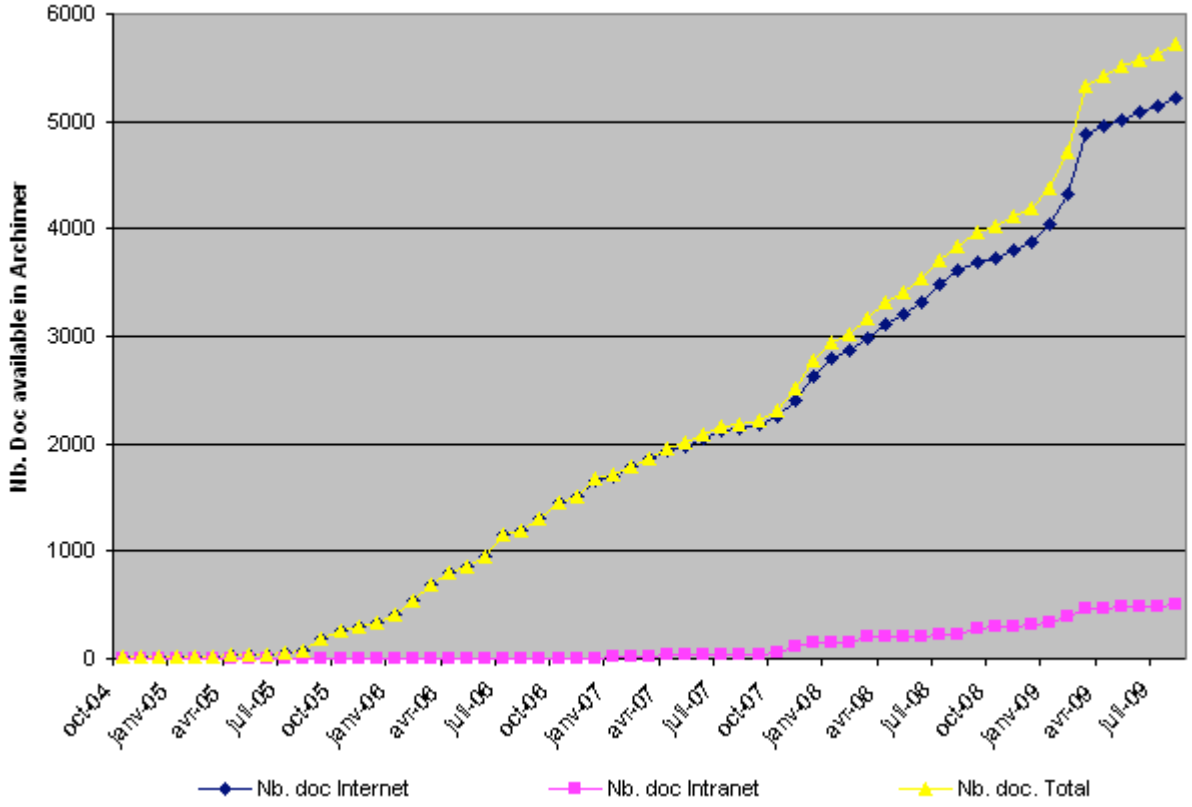

Figure 1: Evolution of the number of documents available in Archimer

	Nb. published before 2000	Nb. published after 2000	**Total**
Reports	484	244	**728**
Theses	37	155	**192**
Publications	1615	1651	**3,266**
Proceedings	901	247	**1,148**
Total	**3,069**	**2,316**	**5,385**

Table 1: Distribution of the documents available in Archimer by document type

Subject	**Nb. of documents**
Biology	2,732
Aquaculture	1,603
Ecology	948
Fishing	896
Pollution	396
Economy	317
Engineering	370
Physics	446
Geology	585
Chemistry	190
Mathematics – Computer science	107
Geography-Land settlement	118
Climatology-Meteorology	99
Law	32

Table 2: Distribution of the documents available in Archimer by subject

Recorded Uses

Access Paths to the Documents Recorded in Archimer :
Different paths are available to access the 5,400 documents accessible in full text (see Figure 2-2) in Archimer. More and more Internet users (see Figure 2-4) know Archimer. When looking for a document, they go directly to Archimer's home page (see Figure 2-1) and use the different search functions offered by this tool.

Users can also access the full text of the documents available in Archimer through standard search engines (e.g.: Google, Bing, Yahoo, …) (see Figure 2-6). In some cases, it is impossible to index the full text directly. This can happen if the files are too heavy, if they are extraction protected or if the PDF files are corrupted. In order to give these files some visibility, we publish, for each of them, a static Web page displaying all the bibliographical information (titles, abstracts, authors, …) and a link to the full text of the document.

A share of the users interested in the documents they found via Google bounces on the Archimer Website (see Figure 2-1) from which they can then discover the entire Ifremer production and, as a first step, the latest publications.

In addition to standard search engines, all the documents archived in Archimer are referenced in a series of harvesters: Oaister, BASE, Avano (see Figure 2-7)…

In a less systematic way, many documents available in Archimer benefit from backlinks (see Figure 2-8) from quotations in other works, from the ASFA base and from library catalogues (for theses and reports).

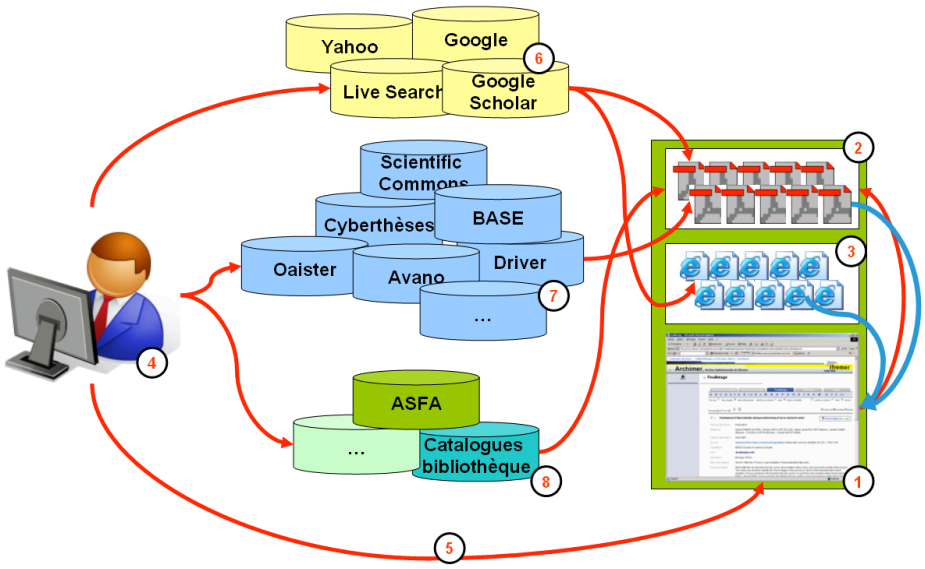

Figure 2: Access paths to the documents recorded in Archimer

An Access Path of Choice to the Documents : Google

Among all the access paths to the full text of the documents, one prevails neatly over the others: Google. This search engine is the source of more than 80 % of the full text document downloads, as shown in the table below. Table 3 presents the ranking of the main access paths followed by Internet users to access PDF files available in Archimer:

Source name	Downloads (in percent)
Google	80,98 %
Google Scholar	4,29 %
Archimer	3,87 %
biblioteca.universia.net	1,55 %
Yahoo	1,51 %
Ifremer search	1,32 %
Bing	1,20 %
…	…

Table 3: Ranking of the main access paths to the documents available in Archimer

A document which full text is indexed by Google will be, on average, ten times more downloaded than a document which is not indexed by this search engine. Consequently, the consultation of a document is not only linked to its interest, but also to a number of technical criteria combined by Google to determine its position in its search lists. The criteria applied by Google are not all known and can vary through time, but we may assume that the following criteria can explain some differences in the consultation of documents:

- **Correspondence between search terms and the words contained in the document.** It is the basic criterion. It most probably explains the fact that heavy documents (like theses) are more consulted than the others. They indeed contain more words susceptible of matching the users' search criteria. However, there is a limit to this rule: documents over 10 MB are not indexed by Google. In this specific case, larger documents are less visible than documents with only a few pages of text. It is also to notice that all words do not have the same importance. For example, the words of the title are more important than the words of the text: if a word contained in the title of a document matches a user's research, this document will have more chance to appear at the top of the Google results than if this word appears at the bottom of the full text.

- **Popularity**: Google chooses among pages of equivalent relevance based on their popularity. To evaluate the popularity of a document on the Web, Google counts the number of backlinks pointing to it. However, all backlinks do not have the same importance. A backlink from a very popular page is more important than a backlink from a less popular page.

- **Originality**: The most consulted documents in Archimer are those which are only accessible through this Website. The relatively poor consultation rate of international publications must be linked to this phenomenon since these documents are also available on their editors' Websites.

Figure 3, below, exemplifies our dependence on Google. "L'élevage de la crevette tropicale d'eau douce" (Fresh water tropical shrimp farming) is a book which is edited by Ifremer and is now out of print. The Edition department enabled us to digitalize it and to publish it freely in Archimer. At first, Google indexed the full text of this document which size is over 26 MB. Thanks to this indexation, this document was one of the most consulted of the Website with more than 300 consultations per month on average. In summer 2007, Google changed its indexation policy and deleted all the documents over 10 MB from its index. As a consequence, consultations for this document dropped by 90 %.

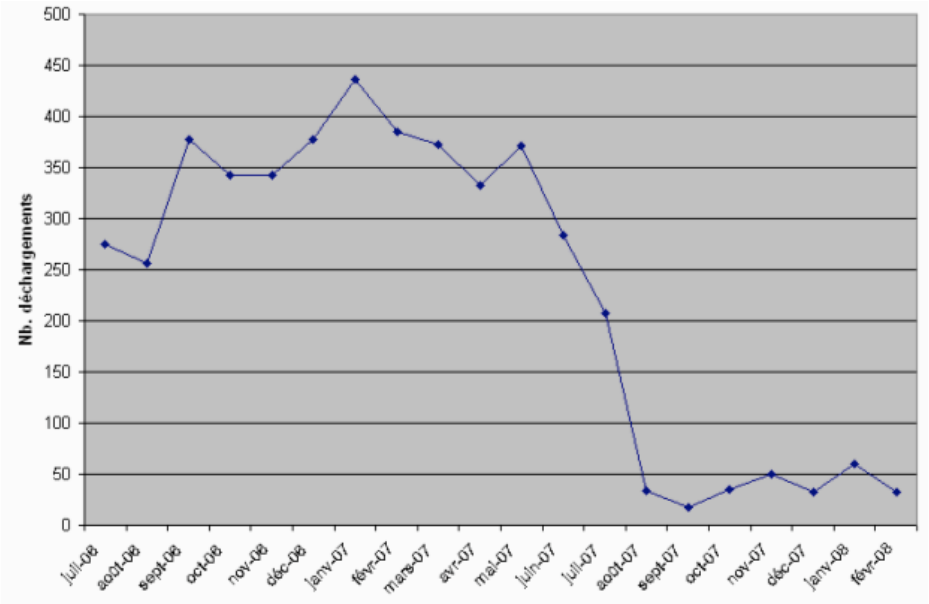

Figure 3: Exemplification of our dependence on Google: Evolution of the number of downloads of the work "L'élevage de la crevette tropicale d'eau douce."

As a complement to Table 3, Table 4 presents the ranking of the OAI harvesters used to download PDF files in June 2009. The use of these harvesters remains marginal

compared to that of Google. This weakness is increased by a difference among harvesters explained by the implementation by the first ranked ones of some indexing systems aimed at standard search engines. As an example, Scientific Commons re-compiles the data harvested in different Archives to create HTML pages indexed by Google (one page per author for example). Thus, Google can be given credit for most downloads from the first harvesters of this list.

Harvester name	**Downloads (in percent)**
Biblioteca Universia Net	1,55 %
Scientific Commons	0,38 %
Avano	0,21 %
Oaister	0,02 %
Driver	0,01 %

Table 4: Ranking of the OAI harvesters used to download documents available in Archimer in June 2009

Evolution of the Number of Connections and Downloads:
Figure 4 presents the evolution of the number of downloads of full text documents outside of Ifremer (downloads by the Ifremer staff are not taken into account). Note the slowdown of this increase, monitored over the October 2007- September 2009 period.

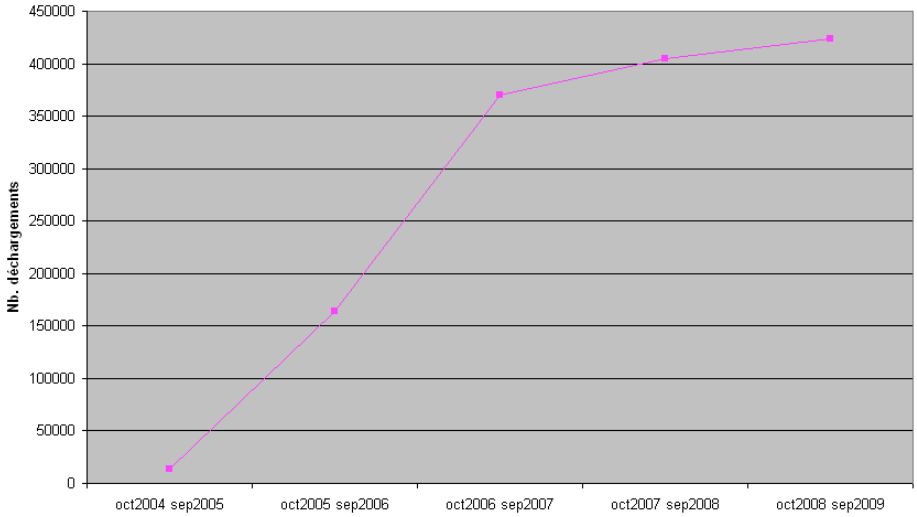

Figure 4: Evolution of the number of downloads of documents recorded in Archimer

Evolution of the number of connections:

Figure 5 presents the number of connections to Archimer's homepage (http://www.ifremer.fr/docelec/). Quite logically, the number of connections increases along with the number of documents available in Archimer, as most of these documents are indexed by Google. If an Internet user finds, via Google, a document he is interested in, he will most probably bounce to Archimer to continue his research.

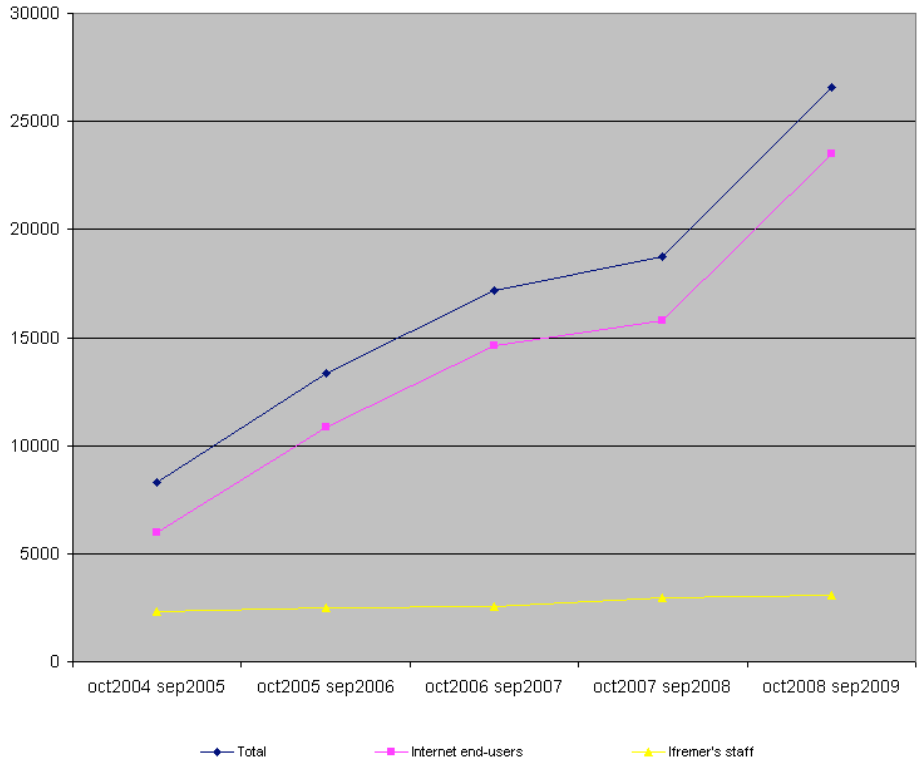

Figure 5: Evolution of the number of connections to the Archimer Website

A Matter of Concern: the Slowdown of the Number of Downloads Since 2008:

Over the October 2005-November 2007 period, the number of downloads was increasing proportionally to the number of documents in Archimer. But since the beginning of 2008, this increase tends to slowdown (see Figure 4) while the amount of documents recorded in Archimer continues to increase steadily. Consequently, it is the average number of downloads per document which has been decreasing since 2008. This drop in the number of consultations seems to affect especially theses and reports although they remain the

most downloaded documents (see Table 5). Moreover, the documents which have been recorded recently are less consulted than those recorded over the past years. Figure 6 illustrates perfectly this phenomenon.

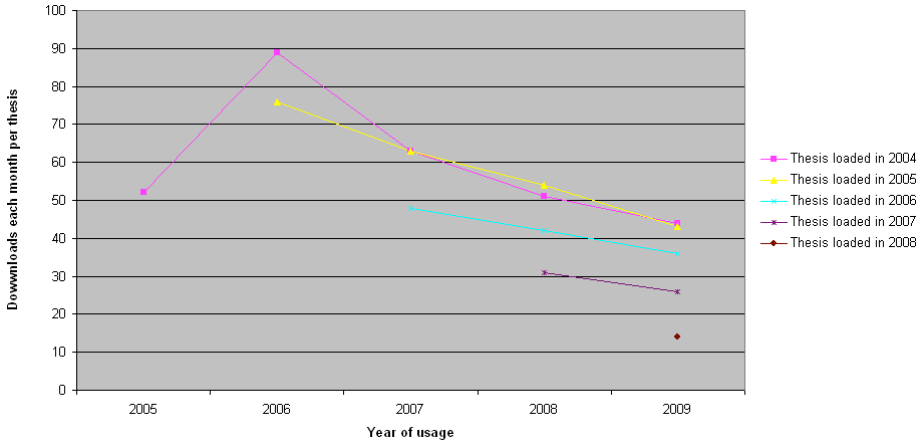

Figure 6: Evolution of thesis consultation according to their year of recording

Explanations to this slowdown are not obvious but we can assume that they are partly linked to the following reasons:

- The gradual desindexation by Google, since summer 2007, of the documents over 10 MB deprived Archimer from a number of successful documents,
- The increasing number of scientific documents accessible freely on the Internet and thus their growing lack of visibility,
- A possible part of auto-saturation. Indeed, we record documents of very similar content. It is especially the case of monitoring reports which content is identical from one year to the other. Moreover, some subjects are overrepresented in Archimer. This is the case of all the studies on oysters (breeding, mortality, …) which account for about 25 % of the documents available in Archimer.
- …

Qualitative analysis of the downloads

Table 5 presents a ranking of the downloads by document type. The observed differences are significant. They are linked to the positioning criteria used by Google (see Section 4.2) and especially to those dealing with the originality and the size of the documents.

	Average
Reports	10
Theses	22
Publications published before 2000	9
Publications published after 2000	5
Proceedings	9

Table 5: Distribution by document type of downloads recorded in March 2009

Table 6 presents a ranking of the downloads by subject. Contrary to the ranking by document type, the differences observed among the different subjects are not especially significant. They must be linked to some technical criteria and not to the relative interest of each document. As an example, in the "Physical oceanography" and "Climatology-Meteorology" categories, Archimer proposes mainly international publications, that is to say documents greatly competed with on the Internet.

	No. docs	**No. downloads**	**Av. download per document**
Aquaculture	895	12,247	14
Engineering	226	7,487	33
Biology	1,382	14,434	10
Chemistry	84	1,899	23
Climatology-Meteorology	30	117	4
Law	19	449	24
Ecology	471	5,577	12
Economy	148	2,494	17
Geography-Land settlement	41	584	14
Geology	161	1,291	8
Mathematics-Computer science	87	1,013	12
Pollution	251	3,824	15
Physical oceanography	196	991	5
Fishing	431	7,471	17
History	16	203	13

Table 6: Distribution by subject of downloads recorded in March 2009

Users' locations:
Figure 7 presents the location of about 87 % of the users, outside of Ifremer, who downloaded, in March 2009, one or more documents recorded in Archimer (the location of the remaining 13 % could not be determined).

Half of the documents available in Archimer are written in French. These documents benefit from the greater visibility as they do not suffer from the competition of other references on the Internet (these, reports, old publications). Thus, in March 2009, French users logically accounted for about 46 % of the downloads, followed by users from North African countries (Algeria, Morocco, Tunisia) with 20 % of the downloads.

If we only consider documents written in English, France still accounts for 15 % of the downloads, followed by the USA (13 %), India (5 %), China (4 %)…

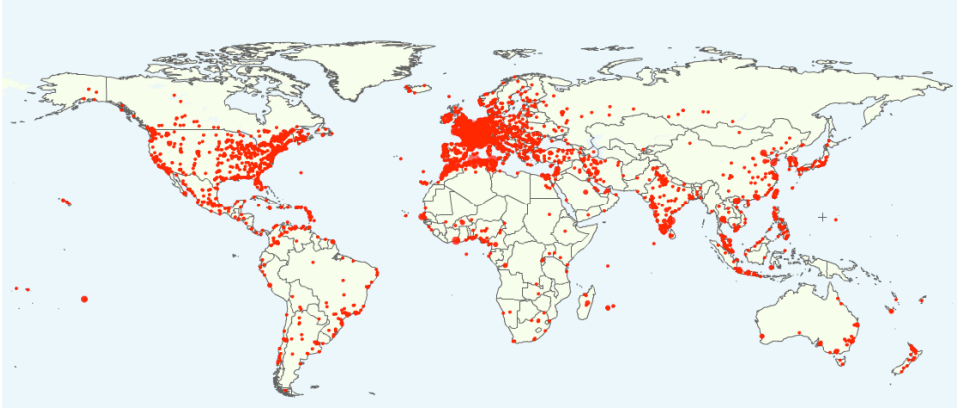

Figure 7: Location of about 87 % of the users, outside of Ifremer, who downloaded, in March 2009, one or more documents recorded in Archimer

Users' Profiles:
The analysis of Archimer users' IP address enables us, in 87 % of cases, to get an idea of the visitors' profile in addition to their location.

In this way, we learn that 20 % of the visitors belong to universities, research bodies or multinational companies.

In about 80 % of cases, we only obtained from the user the name of his/her Internet provider (ex.: Vodafone, Orange, Free, Numéricable…). Consequently, it is difficult to know if they are private individuals, students, or small companies (i.e.: marine professionals).

On the other hand, the analysis of the users' requests can also help us draw their profile. This leads us to think that the documents available in Archimer are mostly consulted by scientists, students and marine professionals (aquaculture professionals for the most part).

Evolution Perspectives:

A new version of Archimer is currently being developed. The main modules should be in operation this spring. This new version will offer a new user interface we hope to be richer and more user-friendly. But most of all, it will integrate new bibliometric functions, including the automatic calculation of production indicators as well as a bibliometric analysis module for Ifremer's scientific production. With these improvements, we hope that Archimer will keep growing as an important tool for the institute's staff.