SET UP AN INSTITUTIONAL REPOSITORY AND AN OAI HARVESTER FOR MARINE AND AQUATIC SCIENCES, AT IFREMER

Fred Merceur
La Pérouse Library
Technopôle Brest-Iroise, BP 70, 29280 Plouzané, France
frederic.merceur@ifremer.fr

Abstract: In August 2005, Ifremer launched Archimer (http://www.ifremer.fr/docelec/), its institutional repository: a full-text database providing free access to publications, theses, conference proceedings, and internal reports. As part of the Open Access movement, this database promotes Ifremer's works on an international scale. A year later Archimer offers more than 1,400 documents of which more than 70% of those publications have been written or co-written by Ifremer since August 2005.

As a supporting step towards the Open Access movement, Ifremer, through the La Perouse Library, also developed Avano (http://www.ifremer.fr/avano/), an OAI harvester for the Marine and Aquatic Sciences. This harvester collects bibliographical data of electronic resources (documentation, images, datasets, videos, audio files) stored in a group of Open Archives and aggregates them in a centralized database. This harvester not only indexes resource references contained in the Marine Science archives of specialized research organizations, but also indexes a selection of resources linked to Marine Science placed in other open archives (ex: ArXiv, Pubmed Central) It is our wish, during this 32nd IAMSLIC conference to present this new project and show its value, especially in the framework of the Aquatic Commons project.

Key-words: Open Access, Institutional Repository, Open Archive, OAI Harvester, Post-publication, Archimer, Avano, Electronic documentation.

Introduction

Since the beginning of the 90's and in order to counteract abusive commercial politics established by some of the scientific publishers, scientific communities created pre-print servers to provide free and immediate access to their work (ex : ArXiv, for Physics and RePec, for Economics).

In 2001, the OAI organization (Open Archive Initiative) formalized an interrogation protocol for those archives. The goal of the OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) protocol is to allow the interoperability of Open Archives. So, if the Archives could not communicate with one another, an end-user would need to interrogate each archive one after the other in order to find a document. Since archive projects are multiplying fast, it is becoming impossible to efficiently conduct a search by using this method.

To simplify the access to documentation available in the archives, the OAI-PMH protocol defines two roles:

- **Data providers** create archives, therefore providing access to resources they enter in them. OAI-PMH compatible archives allow to collect (or harvest) bibliographical data of their resources through a series of standardized commands defined in the OAI-PMH protocol.

- **Service providers** can collect bibliographical data from several archives and gather them in order to create their own database. Therefore, this enables their end-users to interrogate databases corresponding to entire or partial archives. As an example, the Oaister database indexes all of more than 700 archives. Lastly, database records offer hypertext links to full-text documents which remain hosted on the archive servers.

Through Archimer, Ifremer becomes a data provider as a part of the Open Access movement. Through the development of a harvester specializing in Marine Science, Ifremer also becomes a service provider.

Archimer, the Ifremer Institutional Repository archive

Ifremer's interest

Supporting the Open Access movement

Opening an institutional repository brings concrete support to the « Open Access » movement, which progress Ifremer could benefit from in the long run. Rightly so, for several years Ifremer has suffered from subscription raises, just like all other major scientific libraries, established by major scientific publishers and unrelated to inflation. Those raises are forcing it to allot an ever increasing part of its budget to journal subscription contracts to the detriment of other sources of information.

If most international scientific community publications were to become access free on the Web through an Open Archive network, they could become a true alternative to subscriptions offered by scientific publishers. Even without imagining that one day there could be no need for subscriptions, we could be eventually better armed to negotiate our contracts with those scientific publishers on the account of this new data.

Promoting scientific production

Even if free access to all international scientific documentation is a long term goal, setting up an institutional repository at Ifremer should have an immediate effect on its works visibility. Rightly so, several studies show that free access articles are more cited than articles only accessible through Web scientific publishers. Ifremer's free distribution of its publications via Archimer could highly improve their scientific impact.

Creating a new database dedicated to Ocean Science

When the amount of documents available in Archimer will reach a critical level, we hope that the Ifremer staff will see this database not only as a way of promoting its works outside the Institute, but also as a **work base useful for its research. This database should eventually aggregate many documents currently disseminated throughout several servers**. It should also bring access to documents, such as theses which are only accessible through Archimer at the moment.

Renewing relationships between research teams and libraries

Research teams are currently heavily using electronic resources (bibliographical databases, electronic journals…) available to them through libraries. The Institute staff rarely or never goes to library facilities. The staff has access to all resources directly at its desk.

This situation is of course real progress. It allows for the entire Institute staff, no matter where located, to have access to a very large part of the documentation made available by the libraries and to benefit from efficient research tools and documents watch (ex: bibliographical database, automatic research alerts/notification…)

On the other end, this situation tends to isolate librarians from research teams who could underestimate the role librarians play on getting access rights to a selection of information sources. For example, we often come in contact with researchers who think scientific publishers' articles found on the Web are free. Since, the access to those publishers' resources is protected through IP address control; researchers are able to access them without realizing how much work it took to negotiate a subscription contract with for instance a publisher like Elsevier.

Setting up an archive is an opportunity for librarians to strengthen their contacts with researchers when, for example, customized collecting of publications to be recorded in Archimer takes place.

Improving Ifremer Internet Web site's visibility

Documents recorded in the Ifremer Archive are, not only, accessible through the Archimer Web site but also though search engines and the OAI-PMH harvesters search engine.

Archimer end-user statistics show that search engines mainly Google, are the main access points to our documents. They prove the fact that some of the users, who access our documents directly through search engines, continue on their inquiry by going to our Web site Home page. From that page, some look for other documents available via the site. Others continue on browsing by going to the Ifremer Institutional Web site and discover other information related to the Institute.

Documents recorded in Archimer are, consequently, additional new entry points to Ifremer's Institutional Web site. They therefore contribute to an increase in visitors of Ifremer's Web site: an essential communication tool for the Institute.

General Principals

Choosing the development platform

When designing Archimer's system, we made an internal choice using Java, JSP and Oracle technologies, since we originally wanted to reuse part of Archimer's modules in other library projects frameworks, and in particular a Web sites renovation project for catalogue browsing. Furthermore, we hoped to link this new system to other existing computer modules as our Bibliometric database or electronic journals gateway. Focusing on specific development seemed, at the time, to be the best solution to reach our goals.

Web sites included in this project are developed using JSP pages and are carried out through the Ifremer central Apache/Tomcat server. Those technologies were selected for their compatibility with the Ifremer IT department's global policy.

The documents' bibliographical data recorded in Archimer is stored in an Oracle database: data base hosted on a server and mutualised between all of Ifremer's departments. Using Oracle is specifically interesting for Archimer since it allows advanced documentation search functionality integration.

Types of documents recorded

Currently, Archimer is able to record and broadcast theses co-financed by Ifremer, internal reports, conference proceedings and articles published in scientific journals.

In an effort to promote the approval of this project by Ifremer's staff, we first wanted to limit article recording to post-publications only. As of today their free broadcast on the Web is a success as opposed to pre-publications which is sometimes criticized by authors fearing plagiarism where publication quality content is not controlled by peers.

Broadcast format

We chose a PDF format as single format. All documents recorded in Archimer are therefore converted into PDF's; this is so regardless of their word processing tool (Word, Latex…). We selected this format for the following reasons:

- We can be assured of PDF format's permanence, due to its wide use and ongoing publishing of specifications,
- Its implementation is simple, this reduces processing time and document recording in Archimer,
- It is well suited to the electronic broadcast of large files such as publications or theses.

Storing documents

Long term document storing has not been one of our major concerns in this framework. As an example, right from the start we chose not to convert documents into XML/SGML to ensure their permanence. Time spent for such a conversion seemed incompatible with people resources available for the project.

However, when considering the amount of PDF's stored, we assume that if this format was to become obsolete there would be conversion tools available to easily convert those PDF files.

Document recording conditions

Documents are recorded in Archimer by the Institute's librarians who are in charge of:

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

a) Recording theses, conference proceedings or internal reports

For theses, conference proceedings or internal reports, the authors request for us to record their documents.

In order to broadcast this type of document via Archimer, the authors need to provide us with, by email, the bibliographical data necessary to reference their document. They also need to send full-text Word or PDF documents (according to file sizes) by email or on CDROM.

If the author's full-text document includes one or several Word files, we convert and merge them into one PDF file using the Acrobat program before transferring them to Archimer's server.

b) Recording recent publications

Some of the authors tell us which publications they would like to be broadcast on Archimer. In that case, we check which rules have been set up by the publisher of the publication as far as auto-archiving. If the publisher authorizes auto-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:

- Every week, we spot publications written by the Ifremer staff. All of those publications are recorded in Ifremer's Bibliometric Database (see the following chapter).

- We then study each of Ifremer publishers' policy for those publications by using the Sherpa/Romeo Web site. If the author's policy is neither found on his Web site or Sherpa/Romeo, we systematically try to contact the publisher and request the authorization to record his articles in Archimer.

- If the publisher authorizes his own PDF files to be auto-archived (ex: EDP Sciences, The Company of Biologists…) we upload the article's PDF file from the publisher's Web site and record it in Archimer. Most bibliographical data is automatically transferred from the Bibliometric Database to Archimer's. Missing bibliographical data is manually copied from the publisher's Web site. For this last instance, recording is done without having to contact the authors.

- If the publisher authorizes auto-archiving but limits this exemption to his copyright to the author's last draft (the version sent by the author to the publisher: version containing all corrections requested by peers during the proofreading process but which has not been formatted by the publisher), we contact the publication's authors to request that version. If they are able to provide it, we use that version to produce a PDF file matching our broadcast criteria before recording it in Archimer. Two cases can be present :

    - The author submits his publication as one or several Word files (for example, one for text and another for charts and figures) we merge those files before converting them into a single PDF file). We also reformat the first page, not only to standardize our publications, but to also meet publishers' requirements (adding a full and standardized quotation of the publication's quotation, adding a link

to the publisher's Web site, adding explanatory text specific to each publisher…)

- The author submits his publication as a PDF file, we reformat the first page before recording it in Archimer.

An archive linked to other La Pérouse library's documental systems
Figure 1 presents the technical structure of Archimer's main modules and how they link to other library documental systems.

Archimer is linked to the Ifremer Bibliometric Database (see fig. 1/5). The purpose of this database is to list articles, published by the Ifremer staff, in peer-reviewed journals. This database has been developed, within the scope of national indicators set up for the evaluation of French research organizations' scientific production. It is fed through the crossing of data exported from the Current Contents Connect® database and Ifremer's LDAP directory (see fig. 1/6 and 7).

To simplify contacts with the authors for those publications, librarians have access to Archimail (see fig. 1/3) which uses the Bibliometric Database. This tool can generate pre-written and personalized messages according to the publication to be processed. When, in the Bibliometric Database (see fig. 1/5), librarians spot an article published in a journal, which publisher authorizes auto-archiving, they simply need to copy the article's identification and paste it in Archimail. With this identification, Archimail retrieves the data necessary to compose the message from the Bibliometric Database, including the publication's title and all of Ifremer authors' email addresses found in the publication. With this information, Archimail composes a message which can be personalized before sending it automatically to all authors found in the article.

To record a new document in Archimer, librarians connect to a Web site (see fig. 1/4) accessible from Ifremer's Intranet. This Web site offers several Web forms that are specific to the types of documents to be recorded (theses, internal reports, publications…).

The document's bibliographical data is recorded on those forms (title, summary, author…) and will be saved in a database (see fig. 1/1). Those forms also allow recording the full text, as a PDF file, which will be stored on the Ifremer Internet server's disk space (see fig. 1/2).

To record a publication already referenced in the Bibliometric Database (see fig. 1/5), librarians can enter the document's identification in this database. This option allows the automatic transfer of available bibliographical data from the database to entry forms. Librarians can then finalize the recording by typing in missing information (DOI, copyright, full-text).

When recording a publication and to automatically obtain data related to the journal in which the article is published, librarians can also do a search in the Electronic Journals Database (see fig. 1/8). This database contains a cumulative list of all titles to which Ifremer subscribes. This way librarians can enter a few words from the title (ex: aqua* liv*), to find the corresponding title and transfer all of that data to entry forms (journal's URL and complete title, publisher's URL and name).

External end-users can look at available documents via the Archimer Web site (see fig. 1/11). In that case recorded documents records are dynamically built through JSP pages according to end-user requests. Those records provide a link to full-text documents (see fig. 1/2).

Every night, a JAVA program (see fig. 1/9) builds a static HTML file for each newly recorded document (see fig. 1/10). This static file provides the document's record as well as a link to the full-text version (see fig. 1/2). Those static files are built for Internet search engine robots (ex: Google, MSN). This way, document records and full-text versions are directly accessible from those search engines.

Archimer is also OAI-PMH compatible. Harvesters and notably Avano, described further on in this document, (see fig. 1/13), can harvest bibliographical data recorded in Archimer by interrogating its OAI-PMH server (see fig. 1/12). Harvesters will therefore be able to feed their own bibliographical data (see fig. 1/14) using references harvested from several archives and offer from their own interrogation interface an access to Archimer's static records (see fig. 1/10).

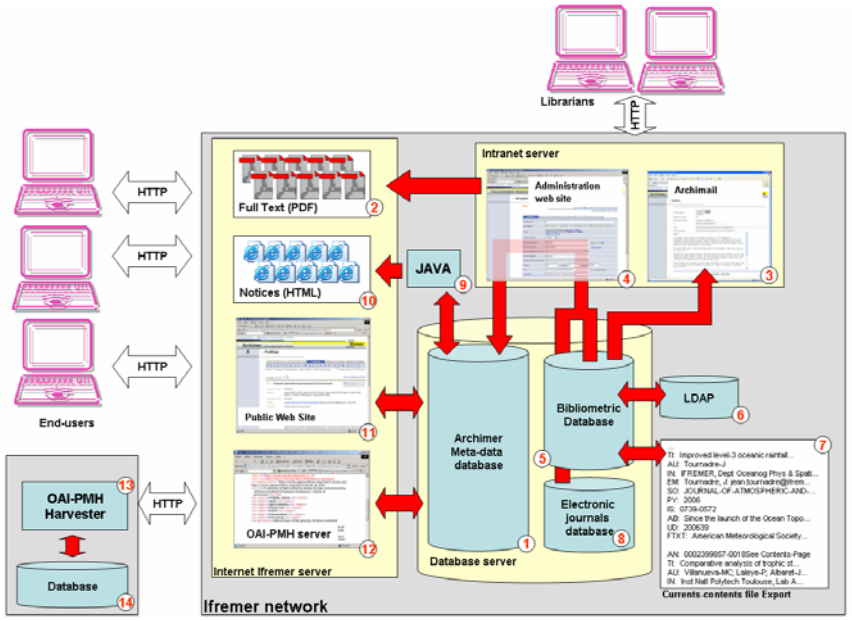


Figure n°1: Archimer system architecture

One year's worth of collecting publications at Ifremer

One year after its opening Archimer offers more than 1,400 documents of which more than 70% of those publications have been written or co-written by Ifremer since August 2005.

From August 1st, 2005 to August 15, 2006, 116 publications with the first Ifremer author have been found in the Current Contents Connect® database. 82 of those 116 publications are already recorded, amounting to about 70%. Those 116 publications can be divided as follows:

- 10 articles were published with publishers who prohibited recording of their publications in an institutional repository (ex: American Meteorological Society, ASLO...),

- 16 articles were published with publishers who authorized auto-archiving of their own PDF files. 8 of those 16 articles are still embargoed. They are recorded but will only be visible in a few months, which should quickly bring the percentage of free access publications to more than **77%**,

- 90 articles were published with publishers who limited auto-archiving rights to publications' last draft. 74 drafts of those 90 articles were collected and recorded.

On a broader perspective, during the same period of time, 257 publications with one or several Ifremer authors, regardless of their position in the authors list, were found in the Current Contents Connect® database. More than 60% of those 257 publications are already access free via Archimer.

Evolution perspectives

Collecting « author versions » publications upon publisher's approval

As of now, few authors spontaneously submit their document to Archimer. We obtained most of the documents currently recorded in Archimer by personally contacting their authors. However, this method has several limitations:

- When we attempt to contact the authors of a publication, they have sometimes already left Ifremer. This can be explained through the fact that more than one year can go by; between the time an article is submitted to a journal to the time it is visible in Current Contents. When a student publishes an article to present his work at the end of his thesis, he often has left Ifremer at the time his article is published and comes up in Current Contents.

- If a publisher only authorizes broadcast of the publication's last draft, at the time we contact the author to get this version, it is sometimes too late due to lost or deleted files.

Therefore we have started to set up systematical collecting for « author versions » as soon as they are accepted by a journal. So, when authors submit their file to us upon publication approval, we can, not only improve Ifremer's publications collecting percentages, but mainly, under reserve of copyright policy compatibility established by the publisher, broadcast « In-Press » publications. We can then be part of speeding the broadcast process of Ifremer's research results, by broadcasting publications several months before they come up on the publisher's Web site.

Spreading the system to other types of documents

As of now, Archimer allows recording and broadcast of theses, post-publications, in-press publications, internal reports, activity reports and conference proceedings. We are planning on integrating other types of documents:

- Patents,
- Posters,
- HDR (Habilitation à Diriger les Recherches ; a French diploma granting a Higher Doctorate degree)

Avano, an OAI harvester for Marine and Aquatic Sciences

Context

One year after the launching of Archimer, La Pérouse Library launched Avano, an OAI harvester specializing in Marine and Aquatic Sciences. This development aims at:

- Continuing on displaying Ifremer's support to the Open Access movement.
- Offering a better visibility of documents placed in Archimer, by aggregating them with papers found in several other archives, in order to create an international database.
- Offering a new centralized tool to the Ocean Sciences community in order to discover data that is currently disseminated throughout many servers.

Functioning principal

Avano is an OAI harvester for Marine and Aquatic Sciences. Therefore, it collects bibliographical data of electronic resources (documentation, images, datasets...) available in a group of Open Archives via the OAI-PMH protocol in order to aggregate them into a centralized database (see fig. 2/3). Its Web interface (see fig. 2/3) offers centralized viewing of resources disseminated throughout several servers.

Avano harvests many archives from Marine Science research institutes. All resources stored in those specialized Marine Science archives are systematically and automatically referenced in Avano. By the end of September 2006, Avano had harvested the following 6 specialized Marine Science archives:

Archive	No. of Doc Availabl e	Description
ArchiMer, Institutional Archive of Ifremer (French Research Institute for Exploitation of the Sea)	1,446	Archimer is the Ifremer Institutional Repository (French Research Institute for the Exploitation of the Sea). It provides free online scientific or technical documents (publications, theses, conference proceedings, etc) in all fields related to oceans (oceanography, aquaculture, fisheries, etc...).
DRS at the National Institute Of Oceanography	418	The National Institute of Oceanography (NIO) in India hosts the Digital Repository Service (DRS) which collects preserves and disseminates institutional publications (journal articles, conference proceedings, technical reports, theses, dissertations,

		etc...).
Marine & Ocean Science ePrints Archive @ Plymouth	1,520	Marine & Ocean Sciences ePrints @ Plymouth is a digital archive providing access to papers produced by the staff of the Marine Biological Association of the United Kingdom, Plymouth Marine Laboratory and the Sir Alister Hardy Foundation for Ocean Science. Marine & Ocean Sciences ePrints @ Plymouth is also an historical archive containing digital copies of early papers from the Journal of the Marine Biological Association of the United Kingdom.
OdinPubAfrica	1,112	Research & Publications in Marine Science in Africa in digital form, including pre-prints, published articles, technical reports, working papers and more.
Repository@NOAA	34	Repository@NOAA (The National Oceanic and Atmospheric Administration) is a searchable database of full-text, online NOAA documents from several selected NOAA programs. The purpose of this project is to establish the feasibility and importance of archiving on a long-term basis full-text NOAA documents in a secure, accessible, and authenticated NOAA electronic repository. The NOAA IR Pilot Project is collaboration between the NOAA Libraries and Information Network, the NOAA Central Library, and the Digital Commons Institutional Repository platform developed by Berkeley Electronic Press.
WHOAS at the MBLWHOI Library	1,190	The Woods Hole Scientific Community Repository, covering Ocean Physics and engineering subjects, Oceanography and Marine Biology

Avano also interrogates a group of open archives not specialized in Marine Science in which are stored, among others, a group of resources linked to Marine and Aquatic resources. For example the ArXiv server specializes in Physical and Mathematical Sciences and contains several publications linked to Oceanography.

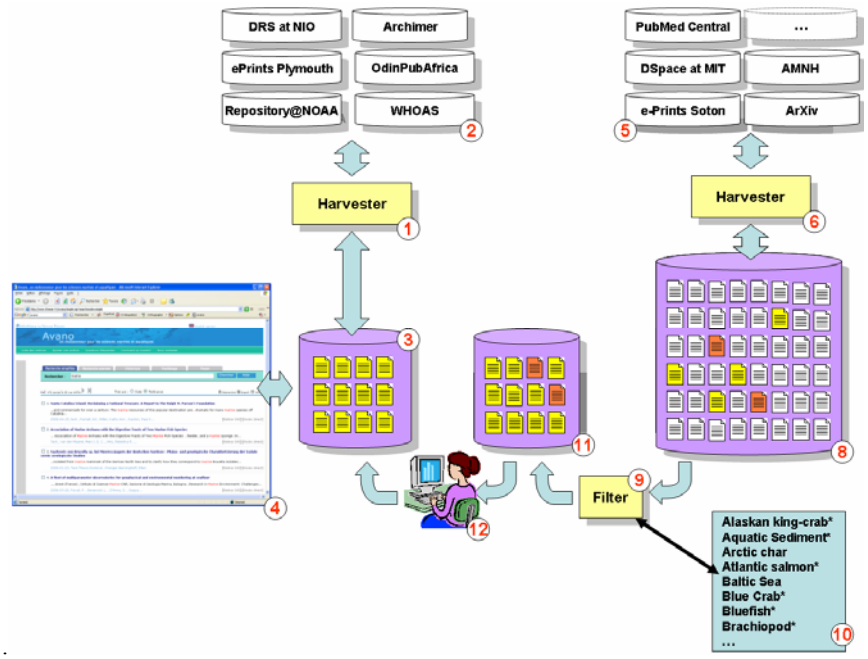
Some of those archives let you isolate documents linked to topics of interest from subsets. In that case, you can automatically isolate resources linked to Marine or Aquatic Sciences and make it viewable to Avano users.

To process archives which are not perfectly categorized within our fields of interest (see fig. 2/5), Avano uploads (see fig. 2/6) all of their records in a temporary database (see fig. 2/8).

Those databases are indexed and an automatic system (see fig. 2/9) isolates records that contain one or several terms linked to Marine or Aquatic Sciences (see fig. 2/10).

Records spotted by this key-word system (see fig. 2/11) are then manually validated by librarians (see fig. 2/12) before they can be visible via Avano. To validate those records, librarians use a Web site (see fig. 3). Key-words found in records are highlighted. This system allows librarians to reject index files when key-words are not related to our field of interest (for example when *Fish* is used for *Fluorescence in situ hybridization*).

By the end of September 2006, this key-word research system allowed us to publish more than 25,000 records isolated within more than 1.5 million records and uploaded from 35 non-Marine Science archives



.

Figure n°2: Avano functioning principal

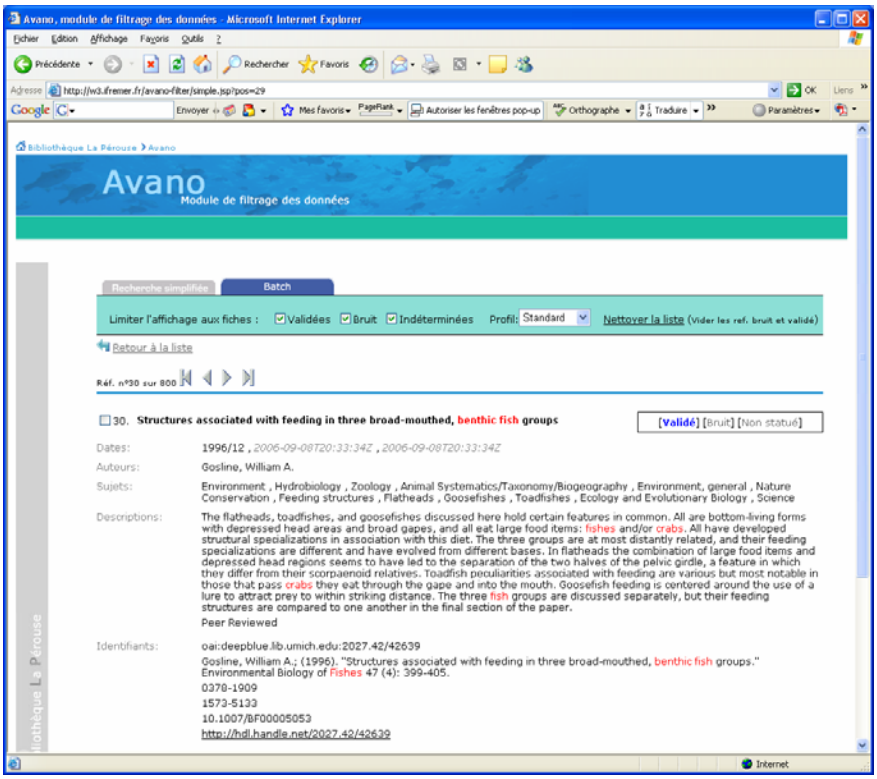


Figure n°3: Avano data filtering module Difficulties encountered while implementing Avano

The difficulties we encountered when setting up this archive are mainly linked to the OAI-PMH protocol limitations:

- **Spotting index files corresponding to a theme in an archive**: This has been the main problem we have been facing. There almost has never been a perfectly matching subset for the fields we wanted to isolate in non-specialized Marine Sciences archives. This limitation led to the development of the key-word spotting system described in the previous chapter.

- **Managing deleted files**: Some archives don't keep track of the files their remove from their database. Those archives are then unable to show the collectors which files have been deleted. In this case, collectors, including Avano, can offer index files pointing to deleted resources. To go around this

problem, Avano will have to completely re-harvest those files on a regular basis to spot potential deletions.

- **Managing doubles**: Several research organizations or universities can record the same electronic resource in their own institutional repository. If Avano collects those archives, it will get descriptive index files of the same topic stored in several places. This can happen if for example a publication is written in collaboration with several institutions. If so, this publication may be archived on those institutions' different servers. Considering our current low auto-archiving rate (environ 15%), displaying doubles in the results list is hardly probable, but this problem should increase in the coming years.

- **Determining publications dates and/or types of resources**: In order to respect the OAI-PMH protocol, archives have to expose their data in the non-qualified Dublin-Core DTD. In this DTD all fields are optional. This optional information trait raises several issues especially for the « date » and « type » fields. When an index file does not have a publication date, it is systematically placed at the end of the list when a user requests sorting his result list by date. Just the same, when a user narrows his search to a range of specific dates, those index files are excluded from the search even if they match the specified search requests.

- **Standardizing the « type » field**: Even if the Dublin Core DTD recommends storing the « type of document » information by using standardized text strings, few archives take this into consideration and still present the information as free text (ex: « publication », « artjournal », « text », « article » are used to describe an article). In Avano, we recommend our users to limit their search to one of several types of resources (documentation, image, set of data, video, audio). To set up this filter we had to implement a standardizing system for this data based on key-word recognition in this character string. This standardizing is therefore imperfect and our filter system may exclude resources from the result list when a user narrows his search to one or several types of specific data.

Evolution perspectives

In the next few months, Avano should be able to harvest more Open Archives; hopefully including new archives developed by members of the IAMSLIC, and therefore would be able to offer a greater number of records to its users.

Furthermore, we may consider also harvesting the Private Publishers catalogue. As of today, two publishers (« High Wire Press » and « The University of Chicago Press Journals Division ») already show their publications with OAI-PMH. If other publishers also adopt the OAI-PMH protocol, we may consider integrating a selection of their records, which full-texts would remain accessible through subscription, allowing users to filter and aggregate them with papers that are free through the Open Archive.

Therefore, Avano would soon be able to provide a more complete view of international research in the Marine and Aquatic Sciences fields.

Collaboration proposition with IAMSLIC

When launching Avano we were pleased to see several IAMSLIC colleagues were interested in this system, among them the initiators of the « Aquatic Commons » project. As a matter of fact we hope that Avano can become a part of that project. In this perspective, we hope to propose to the members of the « Aquatic Commons », even to other IAMSLIC colleagues, joining us for the implementation of this system and in particular for the selection of records originating from non-specialized archives in Marine and Aquatic Sciences.

REFERENCES

Documentation:

Le protocole OAI et ses usages en bibliothèque. François NAWROCKI - Ministère de la culture et de la communication. [28 January 2005] http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm

Archimer, ou la mise en place d'une Archive Institutionnelle à l'Ifremer. Fred Merceur. [23 November 2005] http://www.ifremer.fr/docelec/doc/2005/rapport-657.pdf

Web sites:

Open Archive Initiative site
http://www.openarchives.org/

Oaister
http://oaister.umdl.umich.edu/o/oaister/

OAI explorer
http://re.cs.uct.ac.za/