33rd IAMSLIC Annual Conference
Changes on the Horizon
October 7-11, 2007
Sarasota, Florida, USA

Fred Merceur
frederic.merceur@ifremer.fr
Ifremer – Bibliothèque La Pérouse
BP 70, 29280 Plouzané, France
V1.0

Avano, assessment of one year management of a thematic OAI-PMH harvester

Summary: In September 2006, the La Pérouse Library launched Avano, an OAI harvester for marine and aquatic sciences. Today, Avano provides centralised access to over 100.000 references, the majority of which are freely accessed as full text documents. They were harvested from more than 150 Open Archives.

The aim of this document is to review the functioning principles of this thematic harvester, a year after it was first launched. It also assess the main difficulties met during this first year of management. Finally, this document gives us the oportunity to consider a few solutions to improve the quality of Avano's services.

Key-words: Open Access, OAI-PMH Protocol, Open Archives, Institutional Repository, Harvester.

Table of contents

1. Introduction 2

2. Avano, an OAI harvester for marine and aquatic sciences 3

2.1. General overview 3
2.2. Functioning principles 3
2.3. Assessment of one year functioning 6
2.3.1. A year of harvesting 6
2.3.2. Interrogation statistics 8

3. Difficulties related to the implementation of certain types of repositories and limits of the OAI-PMH protocol 9

3.1. Repository stability 9
3.2. XML stream structure and UTF8 character encoding errors 9
3.3. Harvesting large repositories 9
3.4. Managing doubles 10
3.5. Managing deleted files 10
3.6. Managing the Type field 10
3.7. Managing the Publication date field 11
3.8. Poor quality records 11
3.9. Mixing raw datasets and documentation 12
3.10. Records without free access to the digital object 14
3.11. Thematic harvesting 14

4. Conclusion 14

5. References 16

1. Introduction

Since the beginning of the 90s, some scientific communities have created pre-print servers to provide free and immediate access to their work (ex: ArXiv for physics).

In 2001, the OAI organisation (Open Archive Initiative) formalised a query protocol for those repositories. The goal of the OAI-PMH (Open Archive Protocol for Metadata Harvesting) protocol is to facilitate the interoperability of Open Archives. In cases where the repositories cannot communicate with one another, an end-user would need to interrogate each repository, one after the other in order to find a document. Since repository projects are multiplying fast, it is now impossible to conduct a search efficiently with this method.

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

- **Data providers** create repositories, therefore enabling access to the recorded resources. OAI-PMH compatible repositories allow the collection (or harvesting) of the bibliographical data found in the resources via a series of standardised commands defined in the OAI-PMH protocol.

- **Service providers** (or harvesters), as Avano, can collect bibliographical data from several repositories and compile them in order to create their own database. Therefore, this enables their end-users to interrogate databases corresponding to entire or partial repositories.

2. Avano, an OAI harvester for marine and aquatic sciences

2.1. General overview

Avano is an OAI harvester for marine and aquatic sciences, accessible from the following address: http://www.ifremer.fr/avano/. In September 20007, Avano provided centralised access to more than 100.000 references of electronic resources: a great majority can be freely accessed and contain full text documents.

Avano provides access to resources linked to marine sciences (fishing, aquaculture, marine biology, marine geology, marine economy, oceanography, marine ecology...) as well as resources linked to fresh water resources (lake and river management, wet area restoration, wastewater treatment...)

Avano is partially based on the JAVA version of the Open Archives Initiative Metadata Harvesting Project system developed by the University of Illinois. The filter system presented in the following section and Avano's public interrogation website have both been developed by Ifremer through JSP, JAVA and Oracle technologies.

Today, Avano harvests more than 150 repositories as well as 4 commercial publishers. Avano not only harvests repositories specialized in aquatic sciences but also a whole range of general-interest repositories. In order to isolate records related to aquatic environment in the general-interest repositories, Avano uses a key-word research system described in the following section.

Avano favors repositories providing a large majority of records with a link, free or not, to the digital object. We try and avoid the recording of repositories providing a majority of empty records, without any link to the resource. Of course, this measure is only possible when the repository offers a *Set* allowing only the harvesting of records with a link to the resource.

We also try not to record repositories which provide records pointing at resources located outside their server and in particular, repositories referencing resources collected on the Web.

2.2. Functioning principles

Avano is an OAI harvester: it collects the bibliographical data of electronic resources (documentation, images, datasets...) available in a set of repositories via the OAI-PMH protocol and aggregates them in a centralised database. Its Web interface offers centralised searching and viewing of resource metadata disseminated between different servers.

Avano harvests repositories from different aquatic sciences research institutes. All resources stored in those specialized repositories are systematically and automatically referenced in Avano. In September 2007, Avano harvested 9 repositories specialized in aquatic sciences (see fig. 1). Among all the records found in Avano, almost 19.000 directly come from these 9 repositories:

- Aquatic Commons (Iamslic)
- ArchiMer (Ifremer)
- DRS (National Institute Of Oceanography of India)
- ePic (Alfred Wegener Institute)
- IBSS (Institute of Biology of the Southern Seas)
- Marine & Ocean Science ePrints @ Plymouth
- OceanDocs (Africa and Latin America marine pubublication)
- Plankton*Net (AWI and Roscoff marine station)
- WHOAS (Woods Hole Oceanographic Institution)

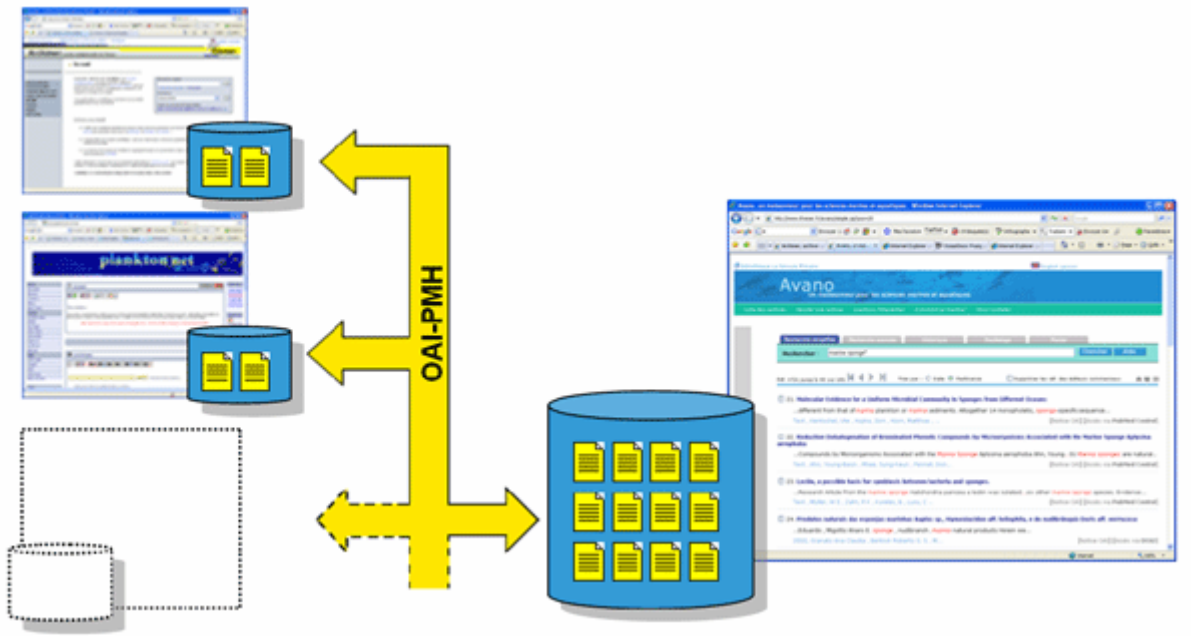

Fig. 1. In September 2007, Avano harvested 9 specialized marine sciences archives. Records provided by these 9 repositories are systematically referenced in Avano.

Avano also interrogates a group of Open Archives not specialized in aquatic sciences which contain relevant resources. This is the case for the PubMed Central server, which specializes in biomedical sciences and life sciences. PubMed Central provides more than 1.000.000 documents of which 15.000 are relevant to Avano's research fields.

In theory, the thematic harvesting of a repository should be made possible by using the *Set* option of the OAI-PMH protocol. Nevertheless, in reality, we have never found any "Marine and Aquatic Sciences" *Set* in any of the harvested repositories. In order to filter those repositories, we have developed a research system based on key-words and key-expressions related to aquatic sciences.

To process repositories that are not perfectly categorized within our fields of interest (see fig. 2.1), Avano uploads all of their records in a temporary database (see fig. 2.2).

Those data are indexed before an automatic system (see fig.2.3) searches for about 30.000 scientific names of aquatic species in the record. For example, if a record contains the character string *Crassostrea gigas* (scientific name of an oyster species), we consider that there is hardly any chance that this name is used in a different context than our field of interest. The record will then be automatically viewable in Avano (see fig. 2.4). This list of 30.000 entries comes from a compilation of a number of lists provided by the FishBase project, the FAO and the National Oceanographic Data Center (NODC). Fred- spell out NODC.

This report is also an opportunity for me to appeal to anybody in possession of a list of scientific names of aquatic species and especially species of algae, fungi, plants, mollusca, gastropoda, insects, birds and mammals. If you are in possession of such a list and if they are not mixed with non-aquatic species, could you please contact me? These lists would be very useful in the automation of our record filtering process.

This first system also searches for over 1.000 names of journals and bulletins specialized in aquatic sciences in the *Source* field of the record. If one of these documents is spotted by Avano in the *Source* field of a record, this record is automatically entered in Avano.

Avano also searches for a range of more general terms and expressions related to the aquatic environment (see fig. 5). For example, Avano searches for the words *fish, marine, fishing, water treatment...* Records spotted by this key-word system (see fig. 2.5) are then manually validated by librarians (see fig. 2.6) before they can be viewed via Avano. To validate those records, librarians use a specific website (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 their fields of interest (for example when *fish* is used for *fluorescence in situ hybridization).*

By the end of September 2007, this filter system based on key-word research allowed us to harvest over 88.000 records found within more than 4.5 million records uploaded from 146 non-aquatic sciences archives.

Of course, this method is far from being ideal:

- This method partially relies on a manual sorting of the records which requires some time (a few minutes per day to filter the new files among the 150 repositories already recorded, plus extra time to process the back-log when new repositories are recorded).
- As we do not spend more than 2 or 3 seconds to either validate a file or not, we may accept a low percentage of records that are not related to Avano's fields of interest.
- We may also miss a low percentage of files, especially when the records are of poor quality (providing no key-words or summary) or when they only provide a text in foreign language.

However, it is the only method we figured that allows us to retrieve about 80% of the records available today in Avano.

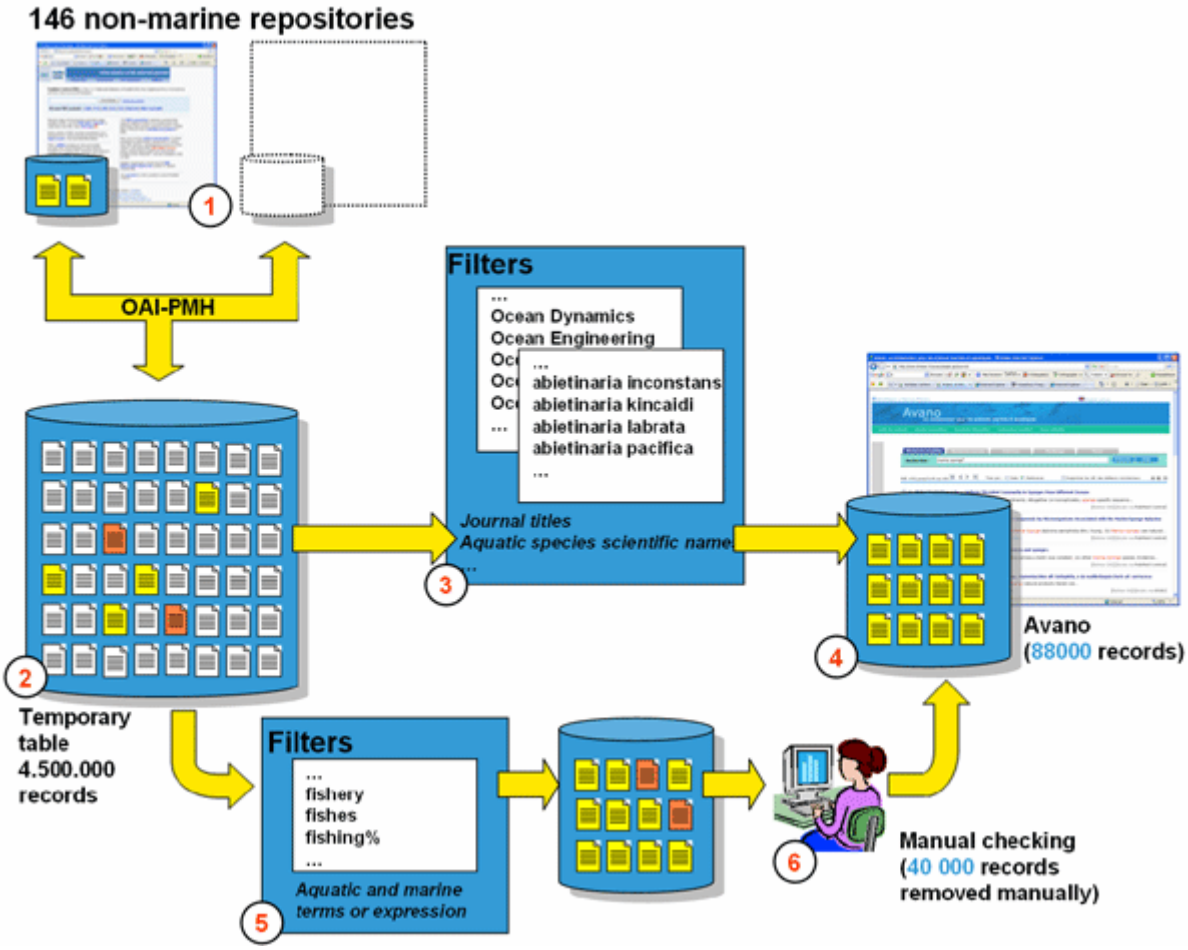

Fig. 2. Avano harvests a whole range of general-interest repositories. Records related to aquatic sciences available in those repositories are isolated by a key-word research system.

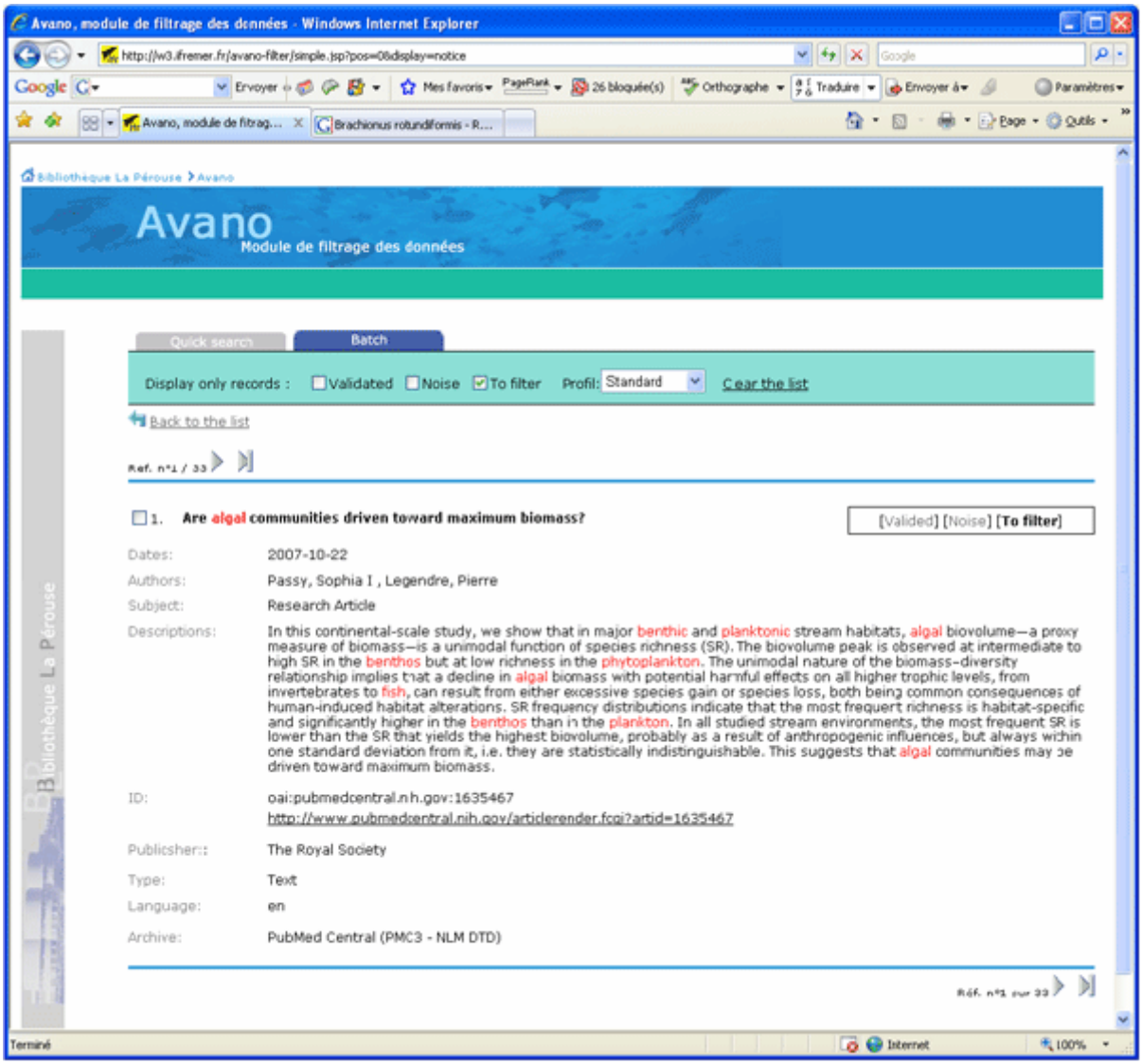

Fig. 3. Manual filtering module for records originating from general-interest repositories and containing one or more terms linked to marine and aquatic sciences.

2.3. Assessment of one year functioning

2.3.1. A year of harvesting

In September 2007, a year after its launching, Avano provided access to more than 107.000 resources originating from more than 150 repositories and 4 commercial publishers. Figure n°4 shows the number of records captured through the year. Most of this progress corresponds to the harvesting of existing repositories and, therefore, to the processing of the backlog. Since most of the existing repositories meeting our criteria have already been recorded, we do not expect similar increases for the oncoming year. In that respect, figure n° 5 is more interesting, as it presents the number of documents available in Avano per year of publication. In this figure, the increasing availability of documents via the OAI-PMH protocol clearly reveals that the Open Access movement has taken off for good, and more dramatically since 2006.

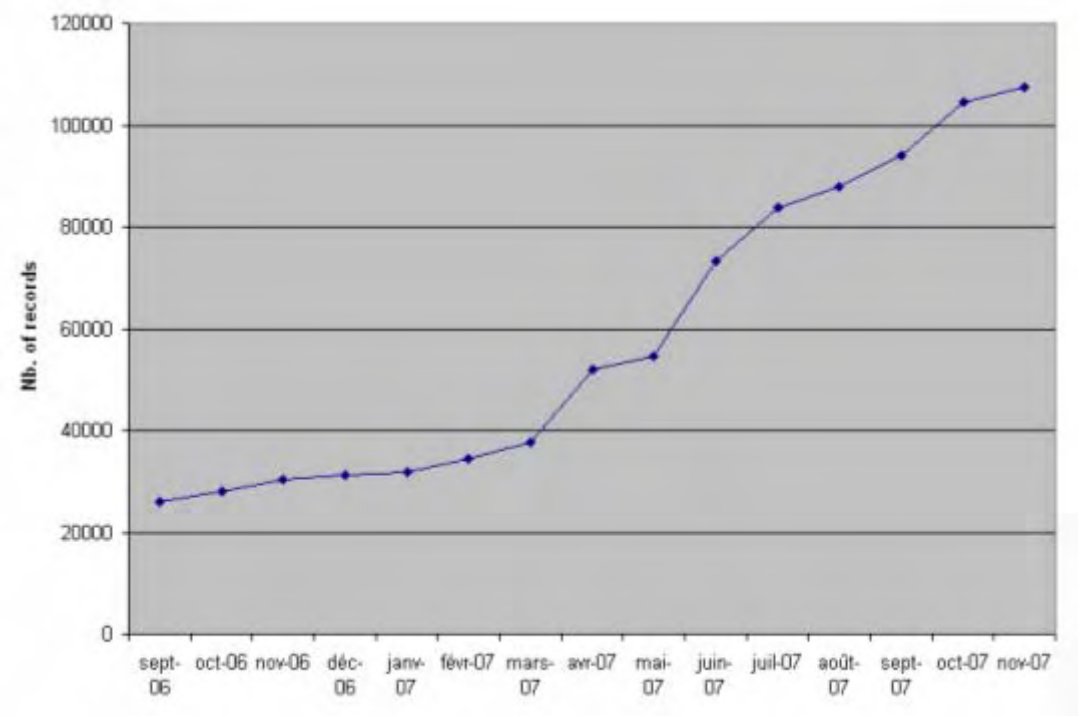

Fig. 4. Increase of the number of records available in Avano since its launching.

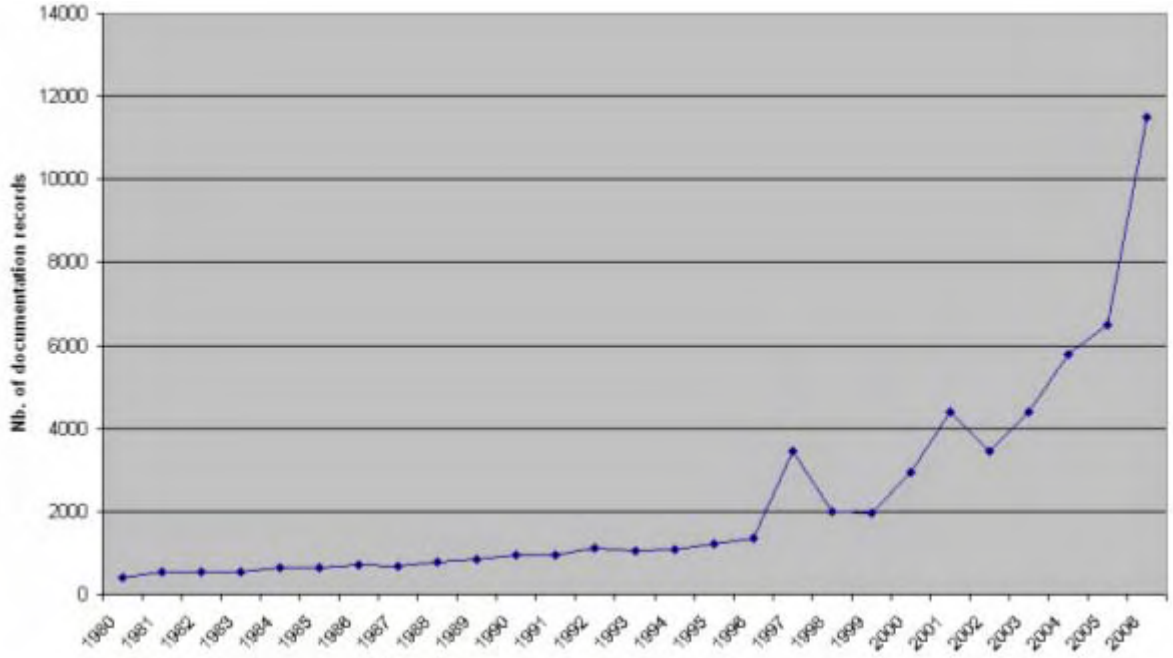

Fig. 5. Number of document found in Avano per year of publication. Only records linked to documents are taken into account (records with no indication of type, images, data files... do not appear in this chart)

2.3.2. Interrogation statistics

Even if it is still low, the number of queries is growing regularly (see fig. 6). The major part of the connexions comes from France (where we have access to more promotion channels) and from the US (see fig. 7).

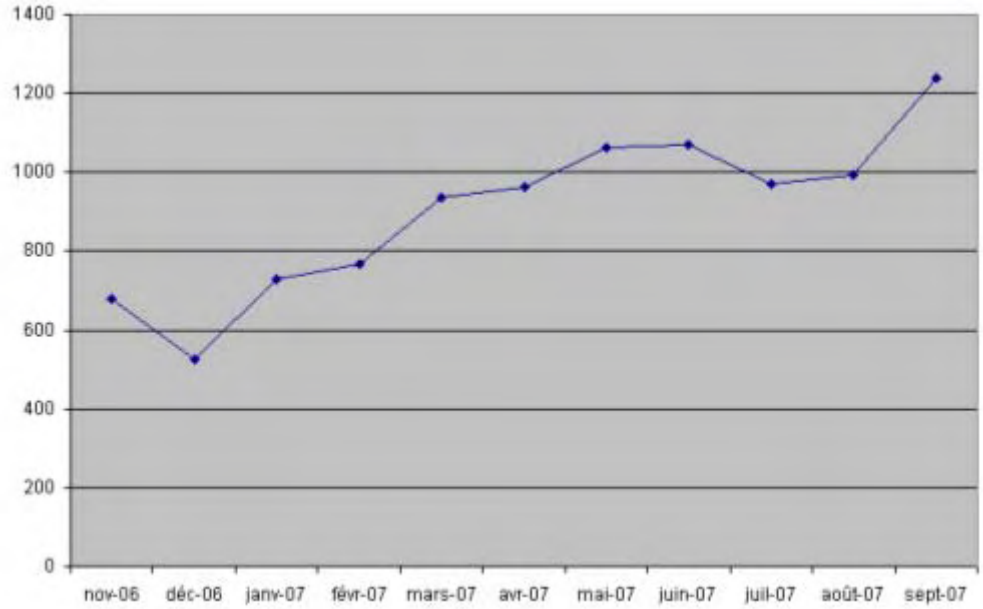

Fig. 6. Progress of the number of connexions to Avano since its launching

Fig. 7. Geographical synthesis, generated by Google analytics, of the connexions to Avano in September 2007. The size and density of the dots is proportional to the number of connexions over the set periode.

3. Difficulties related to the implementation of certain types of repositories and limits of the OAI-PMH protocol

Since its launching, Avano has grown more and more complicated to manage as we've had to face many unexpected technical difficulties. Far from being the automated and autonomous system we had pictured, the management of Avano requires numerous manual interventions. The difficulties we faced during the implementation of this harvester are, for the most part, linked to the limits and the over-permissiveness of the OAI-PMH protocol, as well as problems related to the implementation of some repositories.

Moreover, a certain number of repositories have bad quality data (ex: records without publication date…). And unfortunately, those data are not good for the global quality of the service provided by Avano.

The following sections list the main difficulties faced during Avano's first year of functioning.

3.1. Repository stability

Today, some repositories have stability problems. As a consequence, managing the harvesting of the repositories grew more difficult. The following list is an example of the problems we have to face on a regular basis since the launching of Avano:

- Servers are often inaccessible or return undocumented errors.
- Harvestings are interrupted by http *Timeout* errors, undocumented errors, or without any error message.
- Repositories support the OAI-PMH protocol only partially. As an example, some repositories only support the *GetRecords* method. Others only support the *ListIdentifiers+GetRecords* method. Others do not return the same amount of documents according to the selected method.
- Unnanounced change of servers' URL.

3.2. XML stream structure and UTF8 character encoding errors

Some repositories also transmit records in XML stream that do not comply with the specified DTD. Others return records containing non-compliant UTF8 characters. Those errors can pose problem to some harvesters and particularly to Avano. Indeed, some harvesters use informatic tools that cannot process distorted XML stream. Consequently, those harvesters are incapable of processing a repository containing UTF-8 encoding problems via the *GetRecords* method. Indeed, if a single character is corrupted in an XML stream, harvesters cannot process the records contained in the stream anymore, nor access the following records by retrieving the *ResumptionToken.*

A bypass solution for this type of problem would be to harvest the repositories via the *ListIdentifiers+GetRecord.* In this case, records containing encoding problems are not integrated to the harvester's database, but this method allows, at least, the harvesting of the entire repository. Unfortunately, not all the repositories support the *ListIdentifiers+GetRecord method.*

Another solution consists in contacting the administrator of the corrupt repository and report the errors found in each record. This is probably the most efficient method as administrators can solve the problems really quick, but it necessitates a greater communication network.

3.3. Harvesting large repositories

The initial harvesting of very large repositories and the harvesting of particularly slow repositories can also be a problem (the following harvestings do not pose a problem if those repositories can support an incremental harvesting). The harvesting of those repositories can indeed take a few days, and sometimes more than a week. Since the OAI-PMH protocol does not have a restart point function allowing the harvester to restart the processing from the last record, if any error happens during the harvesting, the process has to be started all over again.

In order to harvest large repositories or slow or unstable repositories, it is sometimes possible to split the process in different time periods, according to the year of update for example. Unfortunately, this is not always possible:

- Some repositories updated all of their records at the same time: in this case, it is impossible to split the process.
- This time-splitting method sometimes brought surprising results as the sum of all the processes did not always match the amount of records contained in the repository.

3.4. Managing duplicates

Too many duplicates in a result list can affect the user's comfort. This is not the main problem harvesters are facing today, but this should increase in the coming years. Today, at least two phenomenons can generate duplicates in the harvesters' databases:

- Several research organisations or universities can record the same electronic resource in their own institutional repository. If Avano harvests those repositories, 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 the server of each institution. Considering the current low auto-archiving rate, especially in life sciences, this phenomenon is not the main cause of the production of duplicates.

- Projects for national or thematic aggregators can pose problem. In some countries, projects of merged intitutional repositories can agregate records from a selection of repositories in a centralised database before displaying them again in OAI-PMH on their own server. As a consequence, records referenced on those servers are displayed twice in OAI-PMH: via the institutional repository and via the centralised database. If the manager of an harvester does not know about the architecture of those national or thematic projects, he may record the two different servers and generate duplicates in his harvester's result lists.

3.5. Managing deleted files

Some repositories do not keep track of the files that have been removed from their database. Those repositories are then unable to show the harvesters which files have been deleted. In this case, harvesters may provide files pointing at resources that do not exist anymore. To go around this problem, harvesters will have to completely re-harvest those repositories on a regular basis in order to spot potential deletions. This requirement can be a problem for large repositories or slow or unstable repositories.

3.6. Managing the Type field

In order to comply with the OAI-PMH protocol, repositories have to expose their data in the non-qualified Dublin Core DTD. In this DTD all fields are optional. Those fields are also *non-qualified*, meaning, for example, that they do not have to correspond to an enclosed value list. This optional and non-formalised information trait raises several issues, especially for the *Type* field.

Indeed, even if the Dublin Core DTD recommends storing the *Type* information by using standardised text strings, few repositories take this into consideration and still present the information as free text (ex: *publication*, *artjournal*, *text*, *article* are used to describe an article). Some harvesters, including Avano, offer their users to limit their search to one or several types of resources (see fig. 8). To set up this filter, harvesters try to standardise the *Type* field using a system based on key-word recognition in this character string. This standardising is therefore imperfect and the filter system may exclude resources from the result list when a user narrows his search to one or several types of specific data. Some informations contained in this *Type* field cannot be standardised. The following list is an example of the *Type* fields harvested by Avano:

- A1
- Article
- 8
- Treball Final de Carrera
- ...

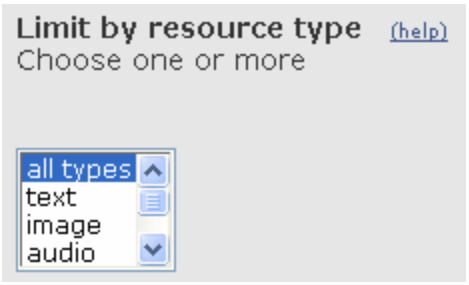

Fig. 8. Option proposed to the users of the Oaister harvester to limit their search to a selection of data types

Even more problematic is the fact that some repositories do not fill in this field. As an example, in September 2007, out of the 107.000 records available in Avano, more than 26.000 did not have a *Type* field. Unless we try to fill in their fields manually (by contacting the repository manager and make sure his repository contains only documents), all of those records are automatically barred from the search space if a user limits is search to one or several selected types.

3.7. Managing the Publication date field

The *Publication date* field poses the same problems as the *Type* field. In September 2007, out of the 107.000 records available in Avano, about 15.000 (a majority of them originating from PubMed Central) did not have a publication date. Furthermore, for a certain amount of records, the *Publication date* field cannot be standardised. This is the case of the following data, harvested by Avano from several repositories:

- 1970-04-00
- 1981.
- Montreal, 2000
- [196-?]
- 2005-92-26
- ….

When a file does not have a publication date or when it cannot be standardised, it is automatically located at the end of the list if the user wants the results to be sorted by date. In the same way, when a user limits his search to a specific period of time (see fig. 9), those files are barred from the search even if they correspond to the specified search.

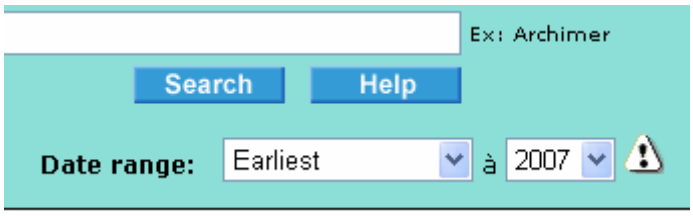

Fig. 9. Available option in Avano's expert search mask to limit a search to several selected data types.

3.8. Poor quality records

Some repositories provide extremely poor records, with only one title and one access to the digital object. If the digital object is a document, its recording in the repository has, at least, the one interest of providing an access point to the search engine robots (ex. Google…). But for the OAI harvesters, those poor records are a real problem. Indeed, a majority of harvesters only index the document records. Moreover, a majority of harvesters provides a default result list sorted by *hit*, that is to say according to the number of occurences of the searched word in the text. In the harvesters, those poor records will then have a low visibility compared to records providing a summary of the document.

3.9. Mixing raw datasets and documentation

Today, a large majority of available repositories mostly provide access to documentation. In the future, repositories could become more diversified, providing, for example, more images, videos, audio files or raw datasets.

As a result, more than 90% of the records available today in Avano are linked to documentation. But Avano also provides access to banks of plankton images (http://planktonnet.eu/), for example. The mix of documentation related records with records linked to images is not a problem and could become, on the contrary, a strong asset for the harvesters.

On the other hand, the agregation of documentation related records with records linked to raw datasets is more problematic. Data in those two domains can indeed be provided with a different granularity. The Pangea server is a good illustration of this problem. This server provides access to hundreds of thousands of raw datasets in the geoscience and environmental domains. Each dataset is described by a record accessible via the OAI-PMH protocol. In this server, thousands of records only differ from one another because of their geographical coordinates. The agregation of this repository with documentation related repositories may overwhelm the result list (see fig. 10).

The Dublin Core description of this type of data offers only little interest for standard harvesters. However, it could be interesting for specialized harvesters, capable of providing their users with a graphic research interface, if the records were presented in a DTD capable of managing graphic coordinates in a standardised way.

10. **Color scan of sediment core MD98-2162**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

11. **Color scan of sediment core MD98-2183**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

12. **Color scan of sediment core MD98-2190**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

13. **Color scan of sediment core MD98-2166**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

14. **Color scan of sediment core MD98-2163**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

15. **Color scan of sediment core MD98-2173**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

16. **Color scan of sediment core MD98-2191**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

17. **Color scan of sediment core MD98-2167**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-25, Dataset , Bassinot, Frank C , Michel, Elisabeth

18. **Color scan of sediment core MD98-2192**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-06-26, Dataset , Bassinot, Frank C , Michel, Elisabeth

19. **Color scan of sediment core MD98-2152**

...Color 700; Color reflectance at 400 nm wave length; Color reflectance at 420 nm wave

2001-03-29, Dataset , Bassinot, Frank C

Fig. 10. The Pangea website provides hundreds of almost identical bibliographical records, differing from one another only because of one information linked, for example, with their geographical coordinates. Thus, this server provides over 1000 almost identical records containing the expression: Color reflectance. If those records are agregated with documentation related records, it will be impossible to find the few documentation records containing the same expression in this thousand of identical records.

3.10. Records without free access to the digital object

The OAI-PMH protocol defines only the sharing process of bibliographical records contained in a group of repositories. As a consequence, some repositories mix records without links to the digital object together with records providing free access to the resource. Others provide records with paying access (ex : BePress) or records with restricted access, for example, for university staff.

In my opinion, this is the major problem harvesters have to face today. There is no indication in the Dublin Core DTD showing the harvesters the degree of accessibility of the objects described in the records. As a consequence, harvesters cannot pass on this information to their users or provide them with the ability to filter empty records or records offering paying access to the resource.

It is my opinion that hiding records with free full text among records with inaccessible full text is not helpful. For lack of time and/or interest, scientists are reluctant to join the Open Access movement and the archiving rate of free access publications stays very low, especially in life sciences. Free and immediate access to documentation is, without doubt, the best way to convince the scientists of the interest of the Open Access movement. And drowning a minority of records providing free access publications in an ocean of records without link to the full text and/or records offering paying access to the documents may not be the best way to promote the Open Access movement.
Again, those records without free access to the full text would not be a problem for the harvesters if the Dublin Core DTD enabled to signify the harvesters the degree of accessibility of the objects described in the records. Harvesters could then provide their users with the possibility of filtering the records without free access to the digital object. But it is still not the case.

3.11. Thematic harvesting

If the thematic harvesting of a repository is considered as possible using the *Set* option of the OAI-PMH protocol, in reality, we have never found any “Marine and Aquatic Science” *set* in any harvested repository. This *Set* is optional in the OAI-PMH protocol and it is, in fact, due to the lack of recommendation, implemented in diverse ways. Some repositories offer a range of thematic *Sets*, sometimes corresponding to a categorization that dates back from their paper collection. Other repositories also offer different *Sets* according to the type of document (publications, internal reports, theses...) or its status (InPress publications, published,...). Finally, other repositories offer some *Sets* allowing the isolation of records providing access to the digital object if the repository also contains empty records.

The implementation of a thematic categorization could be considered among a small community of scientific organisations. But on an international scale, it would be impossible to bring the world scientific community together on a single thematic categorization. This is the reason why we developed this record filtering system based on key-word research. To our mind, this method is the only realistic way of implementing a thematic harvester for all the repositories available worldwide.

4. Conclusion

As we have just seen, even if it is increasing, the number of connexions to Avano today is still relatively low. This can be explained by different factors:

- Every single scientist or student has access to Google/Google Scholar and its billions indexed pages. Scientists also have access to a whole range of Open Archives that have become unavoidable in their domains (ex: ArXiv, PubMed Central…). It is also highly probable that a large majority of scientists, at least in western countries, have access to reference commercial databases (ex: Web Of Science, Scopus,...) covering most of the world scientific production. Compared to Google, which references the full text of a large majority of the documents referenced in Avano, and to the commercial bibliographical databases, harvesters reference only a small part of the world scientific production; mostly because of the low archiving rate of free access publications, especially in life sciences.

- All the harvesters, and they are more and more numerous each day (ex: Oaister, BASE, CyberTheses, Avano, Socolar, Scientific Commons …), share the same audience, providing access to the same documents.

- A harvester does not have a proper content that would provide it with more visibility on the web. The only ways to be known are ordinary promoting operations (mailing lists, referencing on thematic portals). Comparatively, the Archimer website, Ifremer's institutional repository, containing only 2.300 full text documents, is more visited than the Avano website with its 100.000 references. Indeed, each new document recorded in Archimer is indexed by Google. Therefore, it becomes a new access door to Archimer on the web. As a consequence, if a reader finds a document recorded in Archimer via Google (90% of the documents recorded in Archimer are downloaded from Google), and if he is interested in the document, he goes to the Archimer website using the link located in the full text.

- Compared to the standard web search engines (ex: Google), harvesters should provide advanced search options, such as a research by publication date. But, as we mentioned earlier, the bad quality of the bibliographical data provided by some repositories damages the harvesters services.

- Compared to the commercial bibliographical databases, harvesters could have highlighted their free and systematic access to all the digital objects and especially to the documentation. But as we have seen, more and more repositories are drowning a minority of records providing free access publications in an ocean of records without link to the full text and/or records offering paying access to the documents.

So what would harvesters need to find an audience? The following enhancements could be of some help:

- A raise in the archiving rate of the publications in life sciences would help compile a significant amount of free access documents.

- The adoption of the OAI-PMH protocol by more commercial publishers would also allow a quick covering of the largest possible part of the world scientific production.

- The enhancement of the current version of the OAI-PMH protocol would allow an easier harvesting of the repositories and guarantee a better record quality. The enhancement of the OAI-PMH protocol would imply:

    - A restart point system allowing the harvesters to restart a processing that has been interrupted by an error from the last record.
    - Adding normalised and mandatory fields, especially a *Date* and a *Type* field.
    - Adding to the record normalised and mandatory information about the degree of accessibility of the digital object (free, paying, impossible, restricted,...).
    - Adding to the description of the repository information about the involvement of the said repository in a national or thematic agregation system that would reexpose the records in OAI-PMH from a different server. This information would help the managers of the OAI harvesters avoid recording repositories generating duplicates.
    - …

- …

Anyway, today's low connexion rate to Avano is not representative of the success of Open Archives and institutional archives. Other harvesters, and especially those with larger developpement budgets, are certainly more used than Avano. And above all, the documents recorded in those repositories are often massively consulted. They are not consulted via the institutional archives websites, nor via the harvesters but via Google.

Finally, even if the first Open Archive dates back more than 15 years, even if the OAI-PMH protocol is only 6 years old, the Open Access movement has only really emerged since 2006, at least in life sciences. Therefore, we can expect that the archiving rate of free access publications will keep growing to a significant part of the world scientific production. We can also expect the new version of the OAI-PMH protocol to help harvesters provide a better service quality to its user.

Today, the main impact of harvesters may be in the promotion of the Open Access movement, where significant growth is evidenced. Tomorrow, they could be of real benefit in supporting research endeavors throughout the world.

5. References

Timothy W. Cole. What Is OAI-PMH Good For?
http://www.sis.pitt.edu/~egyptdlw/papers/Timothy_Cole.html

Muriel Foulonneau, Friedrich Summann, Jochen Schirrwagen, Paul Walk, Dr. Peter Millington, Laurents Sesink, Maarten Steenhuis, Kasper Løvschall, Franck Falcoz (Feb. 2007). Institutional Repositories Workshop Strand Report Strand title: Open Archives Protocol for Metadata Harvesting
http://www.knowledge-exchange.info/Default.aspx?ID=164

The Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/OAI/openarchivesprotocol.html