Experiences in using the OAI-PMH through the construction of the OA-Hermes Metasearch engine

Egar Arturo García Cárdenas

Main Directorate of Libraries, National Autonomous University of México
Tel: +52 (55) 56223969
egar@dgb.unam.mx

Ana Patricia Gómez Mayén

Main Directorate of Libraries, National Autonomous University of México
Tel: +52 (55) 56223969
pgomez@ifc.unam.mx

Grecia García García

Main Directorate of Libraries, National Autonomous University of México
Tel: +52 (55) 56223969
grecia@ciencias.unam.mx

Alberto Castro Thompson

Main Directorate of Libraries, National Autonomous University of México
Tel: +52 (55) 56223969
acastro@servidor.unam.mx

Abstract:

OAI-PMH has emerged as a more efficient way of facilitating the dissemination of data content. Its success is shown when displaying important information sources that use this protocol, emphasizing: Scielo, ArXiv and PubMed, among others. Nevertheless, its evolution proves a certain degree of arrearage before the current tendencies and demands of the new technological needs, that is to say, the latest technical and methodological characteristics that allow a potential operation of information, are continuously required.

In order to make the new necessity clear on this recognized form of interoperability, the experience gained during the development of the metasearch engine OA-Hermes, which groups different sources of information in a single interface, will be presented. A decrease in time as an achieved factor has to do with the consultation of diverse sources of information, since from a single search, the results of different sources like institutional repositories, digital libraries, and data bases, among others, are semantically integrated.

The technical experience obtained in analyzing the OAI-PMH protocol, led us to reflect on different aspects that have to do with: work methodology, the marketing or fashion of its use, its facility of implantation, or on the information redundancy, just to mention a few.

Key words:

OAI-PMH - OA-Hermes – Meta search engine - Exchange information Protocols - HTTP - Z39.50.

1. Introduction

The first part of the present paper consists in a brief and a step by step description of OA-Hermes; then, to frame the resulted experiences, some analysis will be presented, as well as some final considerations. First, it is necessary to establish that OAI-PMH is an initiative that emerges as an additional option to facilitate the dissemination of digital contents on the Internet. As a collective example of this, there are important and popular open access initiatives that currently offer their services in such a way. However, it is important to mention that the referred protocol presents certain technical delays in its evolution up to this moment, which should be known, considered and taken into account, as well as the advantages offered.

Along the course of this paper, on one hand, aspects such as work methodology, the marketing or fashion of its use, its certain facility of implantation and information redundancy, will be detailed to depth.

On the other hand, it is necessary to mention that there have been some detected disadvantages in the OAI-PMH protocol that will be described along this work, although this is not meant to disqualify it or to suggest stop using it. On the contrary, the intention is to extend the knowledge about it by considering its supporting bases. To achieve this, all the experiences resulted from the construction of OA-Hermes will be set out.

Among the several conflicts that appeared initially, there was the lack of methods to suitably explore the resources from an information source. This resulted in making local copies of all the collections offered by the original source, for their process. This demands the availability of more hardware and software from the collector.

In general terms, the experiences in using the OAI-PMH while constructing the meta search engine OA-HERMES are resumed in a critical perspective, solidly based, that makes emphasis in extending or evolving OAI-PMH as soon as possible for those interested in implementing it. With these aspects in mind, the following sections, that took place in such a way, are to be considered through this paper: integration proposal of Open Access resources; OA-HERMES objectives; OAI-PMH advantages; OA-HERMES conceptual development; detected problems in the construction of OA-HERMES; some OAI-PMH disadvantages; OA-HERMES characteristics, and final considerations.

2. Integration Proposal of Open Resources Access: OA-Hermes

From the beginning of the WWW, outsider information systems that already were on line, have had the tendency to migrate to this environment, bringing within an important increase in communication protocols, information resources, metadata and communication standards, search engines and indexers, e-commerce, e-science and so on, coming to conform, apparently, a parallel world called "e", initial placed before almost any term.

This great increase of the e-world and sources of information has generated a new accessory in our lives, which has forced us to think: "If I can’t find something on the net, it does not exist". It has been clear that, on one hand, it facilitates rich contents and digital services to numerous communities; it also carries several kinds of problems within. That is to say, the third law of Newton is still valid: “For every action there is an equal and opposite reaction”.

Perhaps one of the main framed reactions or problems observed is the heterogeneous exponential growth of information, along with everything involved.

Great efforts are being analyzed and developed to ease and improve the access to information, yet, it is still complicated to access articles, journals, books or another type of digital resources on a certain subject. It is also true that not all the blame can be attributed to the protocol or system. In several occasions, it is the users who are not familiar with the search interface of the resource, nor with the great amount of sources that are available to them. In addition to the previous cases, there are those in which the users carry out repetitive searches in different sites from the Internet, making the retrieval of their information difficult; and the different forms in which the results obtained are displayed by each information source, also becomes an obstacle to be surpassed by the end user.

Taking into account the reasons and problems stated before, OA-Hermes (meta search engine and inter-connector for information sources of open access) emerges with the purpose of facilitating and reducing the time invested in searching and retrieving open access information with an academic validity. Furthermore, OA-HERMES is a tool that favors the integration of collections and repositories of educative institutions in Mexico.

OA-Hermes was gestated within a teamwork of the Main Directorate of Libraries, the Institute of Cellular Physiology and the Institute of Biotechnology of the National Autonomous University of Mexico. When analyzing its potential, the proposal to develop it as a tool available not just for the UNAM, but to open its use and consultation for the academic community of the country and the Internet as well, arises.

By the end of 2004, OA-Hermes needed some financing, so it was presented at CUDI 2004 (University Partnership for the Development of Internet 2 in Mexico). Since the call for papers requested the inclusion of two educative institutions in the country, the University of Colima was invited.

OA-HERMES Objectives

The main objective of OA-Hermes is to group several sources of information in a single interface, in this way, when a user wishes to make a search, he only needs to use the interface that OA-Hermes offers, which directs the search to each one of the sources of information chosen.

In the conception OA-Hermes the following objectives were considered:

- · The incorporation of reliable and high quality sources of information.
- · The access to specialized sources of information many of which are within the Invisible Internet.
- · To take advantage of those open access resources thus enriching the digital libraries of the academic institutions, especially those that have limited economic resources for acquiring or subscribing digital resources.
- · To construct a modular system that would allow its growth and diversification.
- · To favor the visibility of electronic resources produced in Mexico.

OA-Hermes Conceptual Development

OA-Hermes organizes the obtained results from the information sources to be shown to the user later on. Before presenting the data there is a process of semantic integration, in which the metadata of the recovered information are extracted and later unified for their presentation to the user and for the additional processes that could be required.

For the conceptual OA-Hermes design, the following criteria were considered:

1. Extensible
2. Configurable
3. Concurrent Searches, under user's demand.
4. Flexible information management
5. Response time
6. Capacity to be developed by a group.

On the basis of these established criteria, an architecture based on three main components was obtained:

1. The *Nucleus*, which stores and handles the obtained data from the searches in the different information sources. It also directs the searches to the selected sources and provides the results according to the user's demand.

2. The *interface* shows the results to the user. XML is internally handled to display the results, but in order to show the results in different formats, XSL style sheets can be included. In OA-Hermes a style sheet that is used to display the results in HTML is included. This format is the one that is handled by default.

3. *Search engines* are programs that connect themselves to the different information sources and obtain the results from them. These unify the obtained data and send them to the nucleus for their management.

OA-Hermes Characteristics.

OA-Hermes is a proposal aimed to save time for those looking for information on the Web. Instead of going to multiple sites and learning to use their respective interfaces, the user can simply use the single interface that OA-Hermes offers. It is worth mentioning that one of the objectives with which OA-Hermes was conceived, was the simplicity in its internal design and in its user interface, in addition to a low cost of the architecture on which it works.

Making a brief comparison to other search engines, it is important to mention that those used within the Internet, store indexes to organize the information on the part of the Web that is covered. This requires an enormous amount of storage resources. Nevertheless, these search engines do not assume that the included sources of information have their own search mechanisms. OA-Hermes tries to avoid the great amount of information deposits by taking advantage of the storage and search mechanisms that different sources of information offer.

The information sources that are integrated in OA-Hermes handle different communication protocols and formats to display the results. It is frequent to find sources that use Z39.50 as a protocol and MARC or SUTRS as presentation formats. It is also frequent that HTTP is used as a protocol and HTML as a format. Furthermore, OA-Hermes includes the OAI-PMH sources to share their information. The one this paper is focused on is the Z39.50 protocol.

In relation to the operating system environment, OA-Hermes was built with the idea of being a multi-platform, using the Java programming language and the Servlets technology. In order to put it on the Web, Tomcat, in communication with Apache, was used. In the beginning, it was decided to use Perl language for its development but after a careful evaluation, Java was used due to its concurrence handling capacities, IP at several levels, modularization and documentation.

Detected Problems in the construction of OA-HERMES.

Here is a list of the most relevant problems faced while constructing OA-HERMES:

- Incompatibilities and uniformities with Z39.50 protocol (searches, formats).
- Too heterogeneous search mechanisms.
- Open Archives protocol limitations (connection mechanism).
- Heterogeneity in language and characters codification.
- Availability, server's response and connection times.
- Incompatibility among different browsers.

3. Some OAI-PMH Detected Disadvantages

OAI-PMH as we know it is a protocol used to share information through the Web. It is constructed on HTTP by means of commands used by GET or POST methods of the same protocol. These commands are sent to a server which processes the request and sends the results back. Finally, the results are presented in XML and, generally, under the Dublin Core norm.

OAI-PMH offers a series of mechanisms to obtain the resources that a source of information has, and the results can be obtained within a rank of dates (which can be open, that is to say, without specifying initial or final date or both). It is important to mention that the OAI-PMH does not have a sophisticated search mechanism, that is, the results cannot possibly be selected by another type of criteria (author, title, subject, etc., or a combination of these). To summarize, up to this moment, the only way to choose the results from an OAI source is by means of the registration date.

Sometimes the results obtained from an OAI request are too many, thus, these are presented in pages that have a limited number of results (30, 50 or 100 are common values). Each page shows an identifier that corresponds to the following page of the results sequence; in this way, in a new request, the mentioned identifier is included to collect the subsequent data.

Without the intention to totally focus this work in which OAI-PMH technically works, the next sections present, display and document what is considered as the protocol disadvantages, or lack of maturing.

I. Harvester's storage of the obtained results

At this point it is necessary to remember that the systems that provide services with OAI do not provide enough elements for the exploitation of the given data. In order to exploit the information that an OAI source has, it is necessary to store it in a local way, to subsequently, by means of software, manage it as desired. For a small number of results this could not be a significant problem, but when the information source has a higher number of results, special resources are required to maintain and to manipulate this information locally, this is, hardware and software are required for the storage of information and database management respectively, which entails additional costs for those interested in operating this tool.

For example, in the case of Scielo, OA-Hermes extracted approximately 60.000 records, which required a storage space of 120 MB using a data base in MySql.

II. Time allotted to information harvesting and updating

For the extraction of information from the OAI sources, it is necessary to consider the time allotted for the connection and the information transmission, that is, the harvest time. There are two types of harvests: the initial harvest and the updating harvests. The initial one occurs in the first connection that is made to the information source; here, it is expected to extract most of the results that the source has. The updating harvests are used to locally maintain the information stored to the day; these harvests can be programmed in a certain interval of time so as to not highly affect the performance of the service offered.

The time used for the harvests can be considerably long if the information sources have a high number of results. Beginning from ten thousand results, the harvest times can be considered in hours, and if the numbers are higher (millions), we could be talking about days.

For example, the harvest time required to retrieve the approximately 60.000 records from Scielo, was of four hours.

III. Information Redundancy

Among the initial objectives in the creation of OA-Hermes is the avoidance of information redundancy, and that is why the use of the resources provided by the sources for data exploitation is preferred. Nevertheless, there are important OAI sources that must be integrated and, since the OAI-PMH does not provide the operations needed for taking advantage of the information; we are forced to harvest it, that is, to maintain a local copy.

The same case (the creation a local copy) will be present for all those who wish to use the information from an OAI source, which will cause problems like those mentioned previously: the increase in hardware and software requirements; and (for the OAI source users) loss of reliability in data that is not updated, in addition to decreasing the availability in the system when updating the harvests.
Furthermore, the problem is repeated as many times as OAI sources are added to the application that might try to make use of them.

IV. Data Granularity

The results obtained by the OAI-PMH are presented with a series of Dublin Core metadata that are simple and clear enough to be presented to the user. Nevertheless, the Dublin Core elements do not offer enough granularity when additional processes with the harvested information are required.

In OA-Hermes it is sometimes required to recover the full text of the results obtained by an OAI source, with that in mind, it could be necessary to consult data bases that contain the reference to it, or to obtain a URL that leads directly to the full text. The information needed to solve the text (for example the journal, volume and issue) can be in one or more Dublin Core labels, although sometimes the format is not standard, which adds an additional process to enable the extraction of the required information.

OAI-PMH Advantages.

OAI presents advantageous characteristics that without any doubt have contributed to its great success at the present time. These are:

1. The use of standard formats for data interchange.
2. The use and exploitation of XML for the treatment of the extracted information.
3. The use of URLs for the identification of resources which allows taking advantage of the HTTP, the most used and common protocol for information exchange on the Web.
4. The use of Dublin Core to provide a unified platform for the identification and use of metadata, that, although for our aims is not suitable enough, it does simplifies the semantic integration process of the information.
5. Most of contents that use the protocol display an open access modality.

Final considerations

From the OA-Hermes construction point of view, the purpose of the OAI-PMH to help facilitate the efficient dissemination of contents has not been fulfilled entirely, since it breaks the initial ideas of the OA-Hermes project.

During the construction of OA-Hermes, several problems were detected, to which probably, those wishing to operate OAI resources coming from remote sources of information, will have to face.

Another important aspect is to know when the OAI-PMH is really required or needed, because if its use is not analyzed, it is possible to duplicate, triplicate or quadruplicate the contents within a same institution, causing unwanted requirements of hardware, software and personnel that can become excessive maintenance expenses.

The series of problems related to some disadvantages in the OAI-PMH had to do with the immature techniques it uses, although it is recognized as a protocol that has come to stay.

It is also understood that several protocols that are known as standard up to now, have been evolving in the course of the time and they have vanished without leaving a sign.

We state that it is necessary to make extensions that incorporate a greater number of flexible and advanced search mechanisms in order to facilitate and meet the current needs of the institutions using the OAI-PMH protocol. These considerations might improve the use and visibility of the contents which are, in larger number, open access.

References:

Van de Sompel, Herbert ; Lagoze, Carl (ed.) (2004). The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Van de Sompel, Herbert ; Lagoze, Carl; Michael Nelson; Simeon Warner (ed.) (2005). Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/2.0/guidelines.htm

CUDI Reunión de primavera. Abril 2006. http://www.cudi.edu.mx/primavera_2006/programa.htm

CUDI Reunión de otoño. Octubre 2005. http://www.cudi.edu.mx/otono_2005/index.html

CUDI Reunión de primavera. Abril 2005. http://www.cudi.edu.mx/primavera_2005/index.html

Jenn Riley (2005). OAI Best Practices. http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?DigitalTactileResource

Meta-Search Engines. 2005. http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html

Van de Sompel, Herbert (2003) The OAI and OAI-PMH: How did we get here, and where do we go from here?. Delivered at 3rd. Open Archives Forum Workshop, Berlin. Presentation.http://eprints.rclis.org/archive/00001157/02/berl_desompel.pdf