REDALYC OAI – PMH
The Open Archives Initiative Protocol for Metadata Harvesting
(Protocol Version 2.0)

Mtro. Eduardo Aguado López
Universidad Autónoma del Estado de México
Director General de Redalyc
Tel: +52 (722) 2 15 04 94
eal@uaemex.mx

Dra. Rosario Rogel Salazar
Universidad Autónoma del Estado de México
Directora Editorial de Redalyc
Tel: +52 (722) 2 15 04 94
rrs@uaemex.mx

Ing. Arianna Becerril García
Universidad Autónoma del Estado de México
Directora de Sistemas de Redalyc
Tel: +52 (722) 2 15 04 94
aribg@uaemex.mx

Lic. Honorio García Flores
Universidad Autónoma del Estado de México
Coordinador de la Hemeroteca Redalyc
Tel: +52 (722) 2 15 04 94
hgf@politicas.uaemex.mx

Abstract

With the arrival of the new communication systems and the “disappearance” of the borderlines, new challenges emerged to scientific production and spreading. We recognized that great part of that production was unknown and its area of influence is limited. We stated what we already knew, that social sciences were under-represented in databases that determined “the great current of science” and that Latin American scientific journals would hardly enter the legitimized databases. The adoption of the Dublin Core format in Redalyc - that administers a wide database of arbitrated scientific articles of and on Ibero-America it has required to analyze the structure of metadata, in order to implement the protocol OAI-PMH, to fortify the search, recovery and indexing the database through harvesters and search engines over the Internet by making all the articles of its collection available, facing the challenge to give visibility to the Ibero-American scientific production.

Keywords

Dublin Core, OAI-PMH protocol, Redalyc, Ibero-American journals, full text articles, metadata format, scientific journals, arbitrated articles, OAI data provider, OAI software implementation.

Antecedents

Since its creation in October 2002, the “Red de Revistas Científicas de América Latina y El Caribe, España y Portugal (Redalyc)”, set a goal: to give visibility to the scientific production generated in Ibero-America, that is underestimated in the world-wide scene due to different factors like low investment in science and technology, the low participation of Latin American scientists in the "main current of science", measured by the percentage of articles signed by Latin American authors in main databases and the low impact of that production.

The main communication vehicle of academic and scientific means is the publication in scientific journals. In this sense, the Latin American participation of authors or institutions in principal journals or the impact of journals produced in the region allows to know the effect of Latin American science in the world. According to Ricyt (2002), the participation of the Latin American scientists in “the main current of science”, measured by the percentage of articles signed by authors of Latin America in the main databases that register scientific publications was practically null, less than 3% in the important repositories, although there was a growing, for example, 2,7% in the Science Citation Index (SCI).

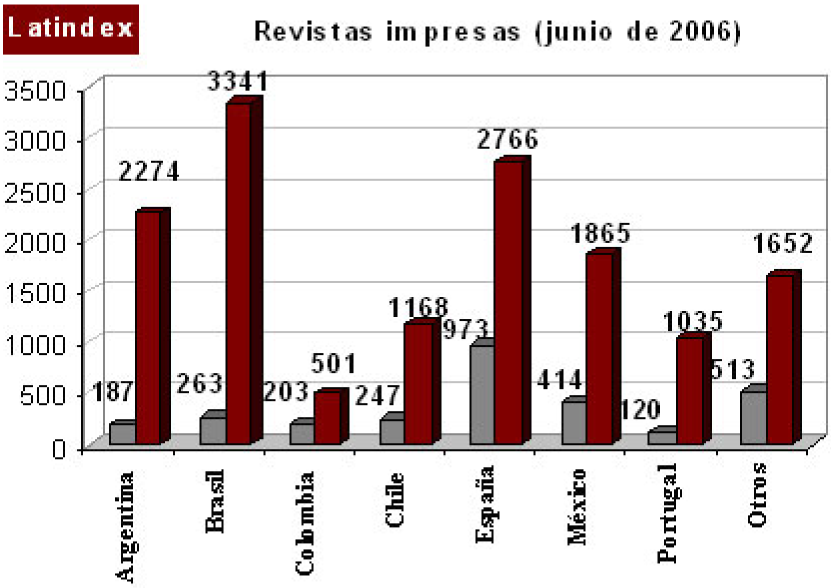
The under-representation of the Latin American production is shown in the composition of the Journal Citation Reports (JCR) of ISI - Thomson Scientific, about 9.000 journals and more than 200 disciplines based on four journal parameters (higher impact, greater frequency of use, greater demand and greater) where social sciences only participate with 1712 journals of that universe, the under-representation is evident. On the other hand, the low weight of the Latin American publications is exhibited when we identified that only 12 Latin American journals - of the 1712- have been able to be indexed in the JCR. If we consider that it is the main base to measure the impact, the conclusion is clear: the Latin American production is not present.

Recent researches show that the cost efficiency in investment and development depends, in a big way, on the possibility that the results are transmitted and consumed by the academic actors in general. The breach between “releases” (published and cited articles) is greater than “entries” (investment and development costs). This polarization demands to modify the levels of “releases” (Dickson, 2004) and sets the need to participate actively in the construction-validation of the science of those countries that are not included in the “great current of science”. The enlargement of the breach shows that increasing the relative weight of the investment in science and development is not enough, but it is equal or more important, the reached effectiveness of the communication of the scientific production.

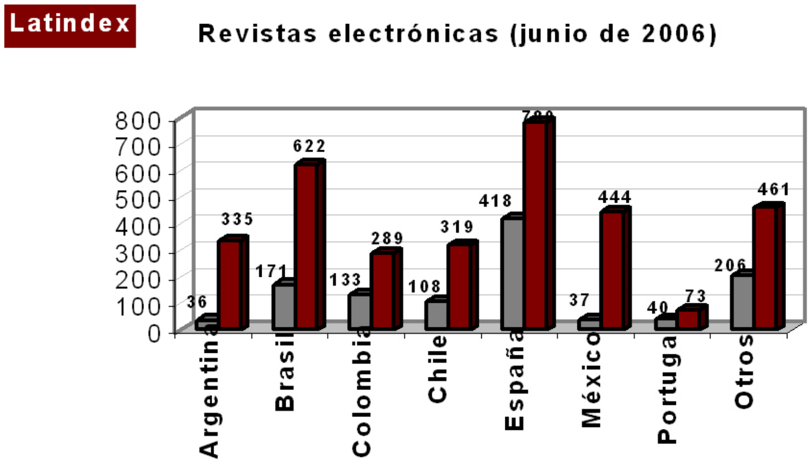
The low efficiency in the communication of the scientific production has taken some specialists to affirm that the major part of the scientific information of the third world countries stays in the penumbra (Gardfield,1999) situation that is turning worst, when it is identified that a researcher coming from the countries in development, is publishing in high international prestige journals, his impact factor is smaller than one of a researcher coming from the countries that control the scientific production (Gibbs, 2001). Therefore, the weakness, the lack of presence and international visibility, the ignorance and the inadequate valuation that journals are experiencing, produced in the Latin American region a central worry of the scientific production’ actors (Cetto, 1998).

An example that the evidences the former premise, can be observed in one of the most important aggregators of content: Oaister, in which, is shown clearly a minimal presence of data providers with Ibero-American collection. There are within the portal, providers like Scielo enriching the collection with more than 50,000 records to the date on its public health database, besides that Dialnet contributes with about 60,000 records, mostly in a referential level. This shows that it doesn't exist a database that contributes with a significantly number of OAI records with link to full text, and focus on the Ibero-American scientific production.

United to this, the lack of technological tools has originated that few journals can offer their contents in an electronic format, even when the efforts have been made to revert this situation: there are still few periodic publications that achieve the access to the electronic format. From the 21,994 registered journals in Latindex just 4,472 belong to electronic journals.

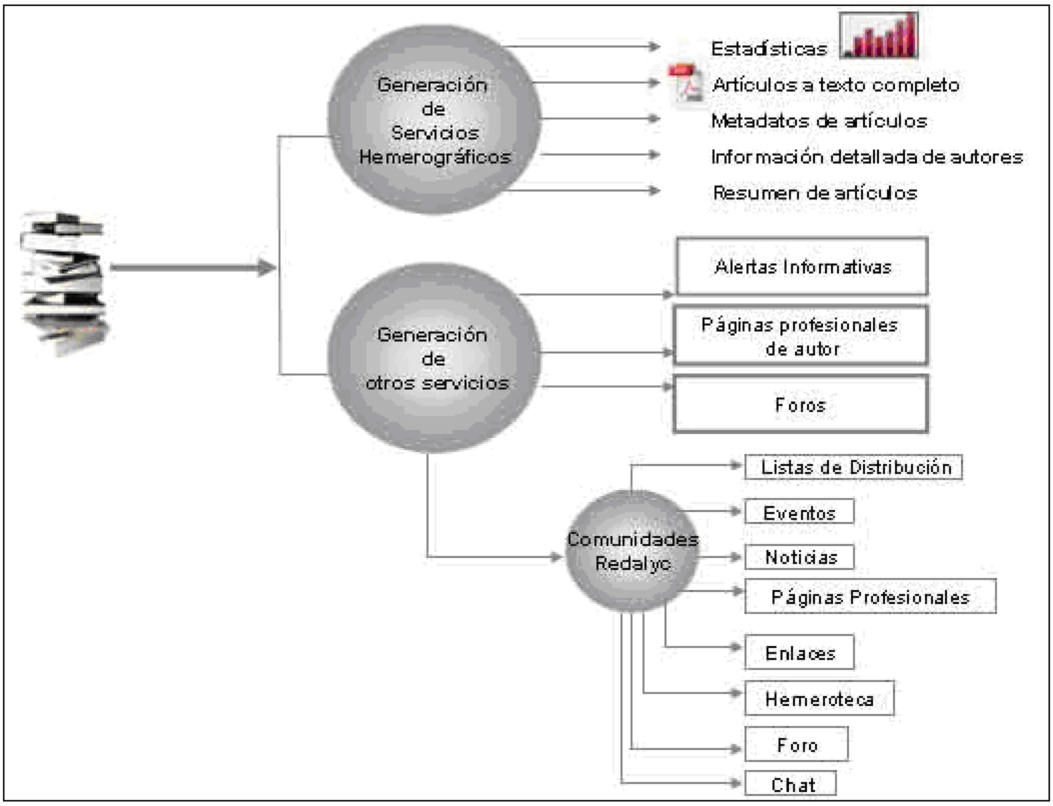


Latindex.
Printed journals (june 2006)



Latindex.
Electronic journals (June 2006)

Taking into consideration that situation Redalyc tries to integrate a greater number of journals than the big databases, keeping a standardized metadata format. This issue proposes a great challenge, to give visibility to the scientific production of the region; as a result Redalyc contributes designing a web site that includes library services of interaction for the publishers, editors, authors and users in general.



Redalyc's architecture.

The different search engines and information interchange protocols search and harvest over metadata to find the documents of their interest, for the metadata description Redalyc uses a code of its own, which includes all the required fields in the description of journals and articles.

Even when the necessary elements are contemplated, at the level of search engines, automatization systems, harvesters and information interchange protocols like the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Z39.50, it's necessary the use of standard formats, such is the case of the OAI, that requires necessarily the metadata format Dublin Core, to be able to use them.

Dublin Core is the most divulgated and cited meta-information format, at least in the strictly librarian scope. It was designed to promote a general purpose standard, simple and descriptive of the electronic resources of any subject (Manso,2002).

Nowadays, the adequacy to interoperability standards is a fundamental requirement for the scientific production spreading that made necessary the adoption of the Dublin Core format, with the intention of being harvested in the OAI official website, besides the distinct search engines like Google Scholar and OA-HERMES and databases, among these we can find mainly the OCLC's WorldCat.

To carry on with this process, it was necessary to analyze the Redalyc's database structure, which in essence works under a relational model. Once identified this elements the next step for the metadata migration process, consisted in the determination of the format to use, choosing to implement the Simple Dublin Core format.

Establishing the requirements and the coincidences by part of Redalyc, to get incorporated to the open archive initiative, the next step was the transformation of the metadata to the Dublin Core, this process implied the programming of a tool that allows to send SQL requests to a relational database, homologate the retrieved records to the Dublin Core format and generate the XML output.

Incorporation of the OAI protocol

The essence of the open access initiative consists in making available to any user, the existing materials on the Web through interoperable repositories to share metadata. This allows that such material can be consulted widely, supporting the basic Redalyc objectives referent to increase the visibility of the scientific production in the Ibero-America countries and collaborating to decrease the notorious disadvantage that exists world-wide in the Internet presence of the scientific collection generated in the region.

The immersion in the open access World, give us the opportunity to participate in the data provider community, contributing to the repositories conformation that will serve as a basis to operate different services to satisfy in a better way, the growing user information demands.

Therefore, the Redalyc OAI – PMH protocol implementation will enable the international access to the library newspaper showing the institutional research outputs and will improve the citation impact in what we call lost science.

Keeping the Redalyc presence in the OAI ambit, important aspect, like it's mentioned in different studies (Steve Lawrence, Brody y Harnad) the number of oa-articles citations increases between a 300% and 500% compare to non oa-articles, we have the expectation to have a definitive raise to the consulted content of at least the triple (read articles-citation correlation). In this way Redalyc not only will become and OAI official data provider (setting the UAEM as one of the three institutions having official participation in the OAI-PMH join to UDLA and the ITESO) but also could compete in the content aggregators positioning like Laoap, OA-Hermes or Oaister, contributing Ibero-American journals to be more known and consulted through the Web and proportionally more cited.

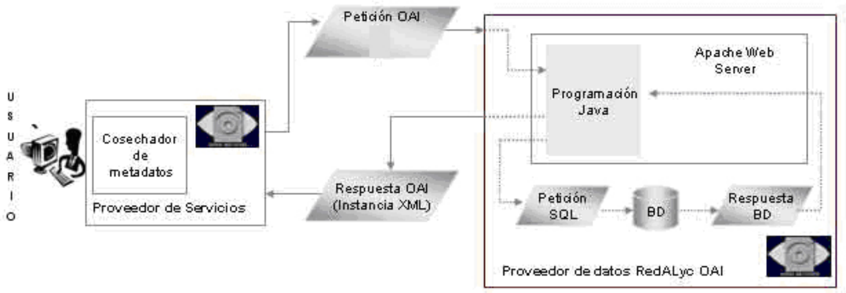
About the protocol implementation it was accomplish by means of the Web development supported by Voai software elaborated by the Universidad de las Américas, Puebla, that helps to generate an OAI metadata server using the Java programming language and Oracle database to produce XML instances. These servers are made to be harvested by other application and the own performance depends on the Web Server, the application code and the operation of the database.

Nevertheless a harvest process is heavy-weighted and server resources demanding. Such performance oscillates between 20 and 200 records by minute of download (Operation of official OAI servers) depending in the amount of metadata that is extracted among other factors,

Redalyc OAI-PMH implementation and software architecture

This software makes a data processing that begins with an OAI-PMH request sent to the server through a metadata harvester and finishes with an XML output instance.

The OAI-PMH gives a simple technical option for Redalyc to make its metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The metadata, that is harvested, might be in any format that is agreed by a community, although unqualified Dublin Core is specified to provide a basic level of interoperability.



Application process.

Protocol implementation details

Implemented Tool

VOAI, software developed by Udlap, Universidad de las Americas, Puebla.

Programming Interface

Java-Servlet 2.3

Record

Redalyc has nowadays more than 27,000 records ready to be available through OAI-PMH. A record has three parts, a header and metadata, both of which are mandatory, and an optional about statement. In Redalyc we cover the following:

Header

identifier: composed of a prefix followed by a Redalyc's id
datestamp: the day since it was available on the web
setSpec: the Redalyc's id of the logical partition of the repository

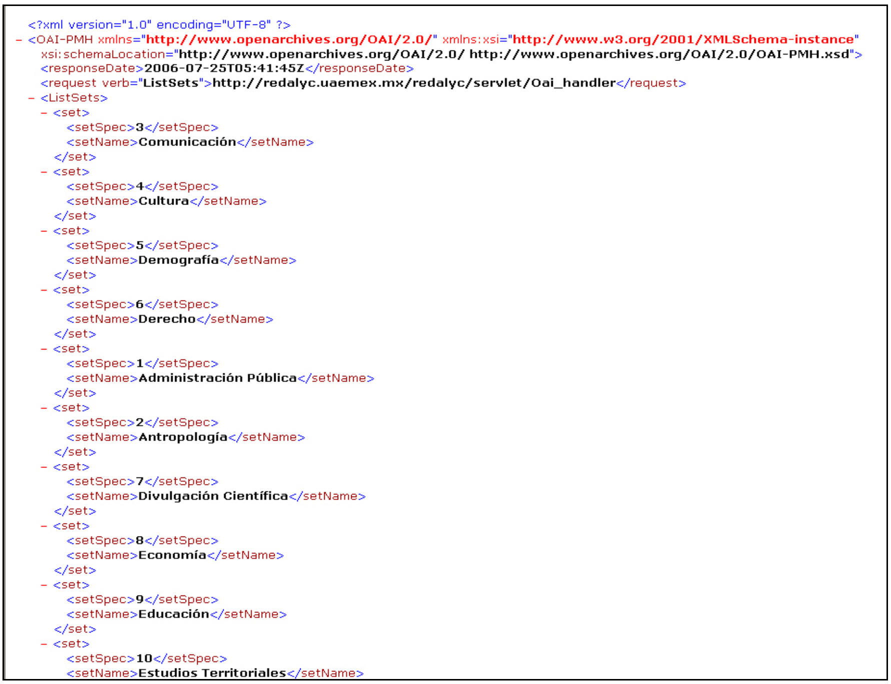
Metadata

Redalyc covers the following DC elements.

<dc:title>
<dc:creator>
<dc:subject>
<dc:contributor>
<dc:publisher>
<dc:date>
<dc:type>
<dc:format>
<dc:identifier>
<dc:relations>
<dc:rights>

Sets

Sets enable a logical partitioning of repositories; the case of Redalyc includes 44 knowledge areas.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
    <responseDate>2006-07-25T05:41:45Z</responseDate>
    <request verb="ListSets">http://redalyc.uaemex.mx/redalyc/servlet/Oai_handler</request>
  - <ListSets>
    - <set>
        <setSpec>3</setSpec>
        <setName>Comunicación</setName>
      </set>
    - <set>
        <setSpec>4</setSpec>
        <setName>Cultura</setName>
      </set>
    - <set>
        <setSpec>5</setSpec>
        <setName>Demografía</setName>
      </set>
    - <set>
        <setSpec>6</setSpec>
        <setName>Derecho</setName>
      </set>
    - <set>
        <setSpec>1</setSpec>
        <setName>Administración Pública</setName>
      </set>
    - <set>
        <setSpec>2</setSpec>
        <setName>Antropología</setName>
      </set>
    - <set>
        <setSpec>7</setSpec>
        <setName>Divulgación Científica</setName>
      </set>
    - <set>
        <setSpec>8</setSpec>
        <setName>Economía</setName>
      </set>
    - <set>
        <setSpec>9</setSpec>
        <setName>Educación</setName>
      </set>
    - <set>
        <setSpec>10</setSpec>
        <setName>Estudios Territoriales</setName>
```

Example: Redaly's ListSets

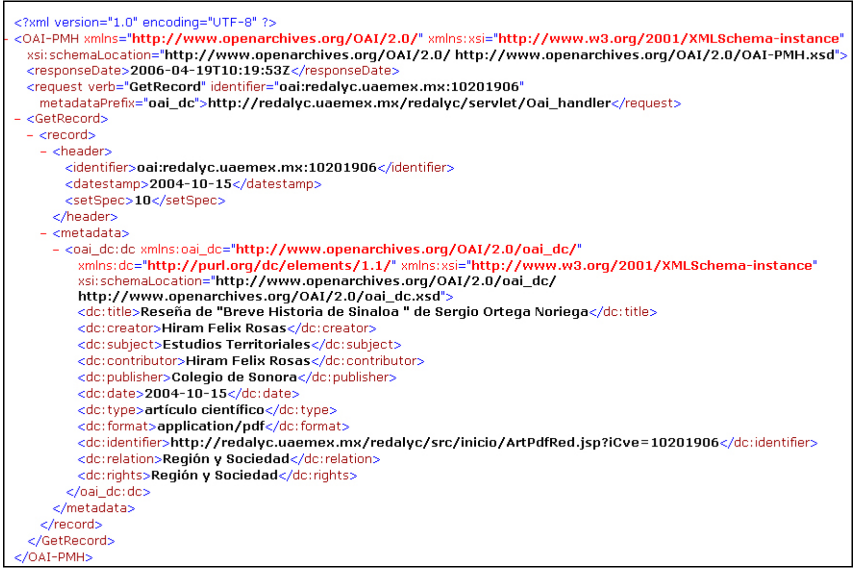
Request format

Requests are submitted using GET/POST methods of HTTP.

Responses

Responses are formatted as HTTP responses. The response format is well-formed XML with markup as follows:

- XML declaration
  (<?xml version=“1.0” encoding=”UTF-8”)
- root element named OAI-PMH with three attributes
  (xmlns, xmlns:xsi, xsi:schemaLocation)
- three child elements
  - responseDate
  - request
  - error (in case of error exception) or element with the name of the OAI-PMH request.



```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2006-04-19T10:19:53Z</responseDate>
   <request verb="GetRecord" identifier="oai:redalyc.uaemex.mx:10201906"
     metadataPrefix="oai_dc">http://redalyc.uaemex.mx/redalyc/servlet/Oai_handler</request>
 - <GetRecord>
   - <record>
     - <header>
         <identifier>oai:redalyc.uaemex.mx:10201906</identifier>
         <datestamp>2004-10-15</datestamp>
         <setSpec>10</setSpec>
       </header>
     - <metadata>
       - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
           xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
           http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
           <dc:title>Reseña de "Breve Historia de Sinaloa " de Sergio Ortega Noriega</dc:title>
           <dc:creator>Hiram Felix Rosas</dc:creator>
           <dc:subject>Estudios Territoriales</dc:subject>
           <dc:contributor>Hiram Felix Rosas</dc:contributor>
           <dc:publisher>Colegio de Sonora</dc:publisher>
           <dc:date>2004-10-15</dc:date>
           <dc:type>artículo científico</dc:type>
           <dc:format>application/pdf</dc:format>
           <dc:identifier>http://redalyc.uaemex.mx/redalyc/src/inicio/ArtPdfRed.jsp?iCve=10201906</dc:identifier>
           <dc:relation>Región y Sociedad</dc:relation>
           <dc:rights>Región y Sociedad</dc:rights>
         </oai_dc:dc>
       </metadata>
     </record>
   </GetRecord>
</OAI-PMH>
```

Example: Redalyc's response.

CONCLUSIONS

With the migration of our metadata to the Dublin Core format, we had the possibility to incorporate Redalyc as a data provider using protocol OAI-PMH, and to increase the visibility, access and diffusion of the scientific production of Ibero-America, besides we are available of being located and indexed of automatic way by different search engines and data bases, contributing to the Redalyc's objective to make scientific resources accessible through the Web.

The internationalization of knowledge is an urgent task of the scientific communities: Redalyc is making its part to reach this goal.

Belong to different communities whose objectives are to offer free access to the electronic resources, will allow to set Redalyc as a safe and reliable option and with effective technological developments to spread and to make accessible the information contained to a greater quantity of users contributing, with it, to give greater presence to the scientific production generated in our region.

References

Brody, T. (2004) The Effect of Open Access on Citation Impact. In: National Policies on Open Access (OA) Provision University Research Output: an International Meeting. Southamptom.

Budapest Open Access Initiative (BOAI). www.soros.org/openaccess/

Cetto, Ana María (2001), "El impacto de las revistas y cómo incrementarlo", en *Seminario CONACYT-UNAM para editores de revistas académicas*, 3-4 de octubre, México.

_______ (1998), "Ciencia y producción científica en América Latina El proyecto Latindex", *Internatl. Microbiol.* 1:181-182, Springer-Verlag, Ibérica.

_______ (1998), "Las revistas científicas como fuentes de bases de datos. Experiencias del Taller de Guadalajara", en el *Taller de obtención de indicadores bibliométricos*, Ricyt-Cindoc, Madrid, 23-25 de febrero.

Dickson, David (2004), "Scientific output: the real 'knowledge divide'", editorial en *SciDevNet*, 19 de julio de 2004.

Dublin Core Metadata Initiative [Consultado 19 de abril de 2006] Disponible en: http://es.dublincore.org/index.shtml

Garfield (1999), www.isinet.com/isi/

Gibbs, W. Wayt, (2001), "Ciencia del tercer mundo", en Eduardo Loría Díaz (ed.), *Viejos y nuevos dilemas de las revistas académicas*, UAEM, México, pp. 101-115.

Harnad, S. and T. Brody, (2004). Comparing the Impact of Open Access (OA) vs. Non-OA articles in the Same Journals. DLib Magazine 10(6)

King, David (2004), "The scientific impacto of nations", *Nature*, vol. 430, 15 julio 2004, en www.nature.com.nature.

Manso Rodríguez, Ramón A, (2002). Aplicación del Formato Dublin Core para la descripción de los recursos en la Biblioteca Virtual del CDICT- UCLV, en Revista Latina de Comunicación Social, número 51, de junio-septiembre de 2002, La Laguna (Tenerife), en la siguiente dirección telemática (URL):http://www.ull.es/publicaciones/latina/2002mansojunio5104.htm

Metainformación - Dublin Core: Elementos del conjunto de metadatas de Dublin Core: Descripción de Referencia. [Consultado 19 de abril de 2006] Disponible en: http://www.rediris.es/metadata/dublin_core_elements.es.html

ISI (2004), www.isinet.com/isi/, consultado en el mes de julio.

Ricyt, (2002), El estado de la ciencia. Principales indicadores de ciencia y tecnología iberoamericanos / interamericanos 2001, Red iberoamericana de indicadores de ciencia y tecnología (Ricyt), Buenos Aires.

II Taller Sobre Publicaciones Científicas en América Latina, conclusiones y recomendaciones (1997), Documento de trabajo, Universidad de Guadalajara, México, 27-29 de noviembre.

http://www.openarchives.org/

http://www.oaforum.org/tutorial/

http://ict.udlap.mx/software/voai_dist/index.htm