Scielo Articles Available to the Agricultural Community Using OAI-PMH in AGRIS AP XML Format

Stefka Kaloyanova, Gustavo Fonseca, Fabio Batalha, Solange Santos, Renato Murasaki, Abel Packer, Steve Katz, and Johannes Keizer

Abstract:

This article reports a new way of getting access to SciELO full text journal articles using AGRIS AP XML format to harvest SciELO metadata and to include them into the AGRIS repository.

It covers the following main steps of the work carried out:

- Selecting about 43 journals with agricultural thematic from the Web site of SciELO;
- Defining the methodology for harvesting;
- Harvesting legacy data and then doing incremental harvesting for the new data from the selected SciELO journal articles for inclusion in the AGRIS XML repository;
- Inclusion of SciELO articles and open access to them through the AGRIS search portal at http://www.fao.org/agris/search/search.do?query=%2Bcenter%3A%28XS%29
- Testing and proposals for improvement and future use of this feature.

We share the methodology used, problems encountered, and the expected benefit. This work proves that semantically rich metadata for agricultural science and research publications based on the "AGRIS Application Profile" from the SciELO repository can be handled by the OAI-PMH protocol. It shows how the selected subset of metadata created with an ISIS application can be harvested through OAI-PMH protocol, which in turn allows for further creation of additional services by giving greater access and visibility to SciELO data in the new AGRIS AP format compared to the used DC format.

The strategy that we adopted was to adapt BIREME's OAI-PMH plug-in for direct generating of AGRIS AP XML from the SciELO application. The existing BIREME OAI-PMH plug-in interface was upgraded to accept and expose metadata using AGRIS AP in addition to the existing DC schema. This approach was elegant but required more time for realization and implementation by BIREME and FAO staff.

Introduction:

SciELO - Scientific Electronic Library Online is a model for cooperative electronic publishing of scientific journals on the Internet, especially conceived to meet the scientific communication needs of developing countries, particularly Latin America and the Caribbean countries.
The SciELO Methodology includes 3 components:
(1)Enabling the electronic publication of complete editions of scientific journals, the organization of searchable bibliographical and full text databases, the preservation of electronic archives and the production of statistical indicators of the scientific literature usage and impact;(2)Application of the SciELO Methodology in operating web sites of collections of electronic journals;(3) Development of partnerships among national and international scientific communication players - aiming at the dissemination, improvement and sustainability of the SciELO Model.

FAO in collaboration with other partners like the CGIAR (Cooperative Group on International Agricultural Research) and GFAR (Global Forum for Agricultural Research) aims to improve global access to agricultural knowledge and information. For this reason some initiatives have been launched to get agreements on data and information exchange standards (http://www.fao.org/aims). One of the important pillars in this collaboration is the AGRIS

Application Profile - a metadata schema with the purpose to facilitate the exchange and harvesting of medium complex, high quality bibliographic data. Compared to the simple DC format usually implemented in the OAI Harvesting mechanisms, the AGRIS AP offers a richer set of metadata and qualifiers including the used vocabularies. The use of the AGRIS AP in harvesting bibliographic data makes it possible to retrieve and trace much better knowledge from these publications.
Within the existing AGRIS network this AGRIS Application profile has been widely introduced and accepted. This made it possible to transfer a highly centralized process of data production into a decentralized but interoperable system. The architecture for the new network has been described (Subirats, I. 2007)[1]) and is the basis for various pilot projects with national networks.
FAO and BIREME have a longstanding collaboration. Some of the SciELO journals are catalogued and searchable through the library catalogue and access to the articles is given through the FAO Virtual Library. Linking SciELO journals into the global agricultural network and providing a possibility for additional search of the selected journal articles through AGRIS search engine is important major step, giving new access and more visibility to the content of SciELO journals.

The Process

SciELO is a center of excellence for online journals in the Life Science area. By checking the list of SciELO journals at http://www.scielo.br/scielo.php?script=sci_alphabetic&lng=en&nrm=iso it was possible to identify 43 journals with relevant subject content to agriculture at large.

The work was then done with the following steps:

1. **Defining the methodology for harvesting**.

Two different approaches were studied for harvesting SciELO on-line scientific journal articles and their further inclusion as AGRIS AP (2) XML files in AGRIS repository:

a) **Using the existing SciELO XML formats**

SciELO metadata are already exposed in simple Dublin Core XML and in a detailed XML format which is used for Pubmed.

A simple example of the script used for ListRecord in order to get DC format metadata for journal with ISSN 0001-3765 can be seen at
http://www.scielo.br/oai/scielo-oai.php?verb=ListRecords&metadataPrefix=oai_dc&set=0001-3765

The possibility to use DC XML and apply XSLT transformation for converting it to AGRIS AP XML was evaluated. Mapping from DC to AGRIS AP was done. Some of the problems we met were related to the lack of some of the mandatory for AGRIS AP elements that were not present in the SciELO Simple Dublin Core XML file. Proposals on how to generate missing AGRIS AP elements were considered.

An example for the Pub Med version of SciELO metadata is shown at http://artigos.scielo.br/S0001-371419980003.xml . This format contains more detailed data. It is used by Pubmed Central and includes also some Google scholar specifications. Mapping for conversion from this format to AGRIS AP XML was done. The AGRIS AP standard conversion and implementation in means of content of the fields was evaluated. Some of our findings were the following:

- author's names were different from AGRIS rules e.g. as Smith, A. John instead of Smith, A. J.
- the publisher, authors' details, and citation articles at the end of the record were available through the links provided.

The proposal for using those two formats was considered not to be a good solution for AGRIS AP harvesting and was not approved.

b) **Direct generating of AGRIS AP XML from the SciELO application**

The alternative solution was to produce direct AGRIS AP XML from the BIREME Database. BIREME decided to use a plug-in different from the OAIAGRIS plug-in developed by FAO (Kaloyanova, S. 2008)[2]). Their strategy was to adapt the existing BIREME OAI-PMH plug-in script (PHP program) by adding a possibility to produce AGRIS AP XML format in addition to DC format. This was at the first stage used for harvesting legacy data. Afterwards an incremental harvesting run by AGRIS harvester for getting only the new data from the selected 43 SciELO journal articles for additional inclusion in the AGRIS AP XML repository will be used. This approach was more elegant but required more time from BIREME and FAO staff.

2. Adaptation of BIREME's OAI-PMH plug-in to produce data in the AGRIS AP XML format

The existing BIREME OAI-PMH plug-in interface was upgraded to accept and expose AGRIS AP in addition to the existing DC format.

AGRIS AP syntax[2] is a more complex, agricultural community specific metadata format, richer than the Simple DC, with mandatory and nested elements that respect and explore a more complex structure of the original metadata for further integration in value added services (Salokhe, G. 2007)[3].

AGRIS AP structure was studied by BIREME's staff. FAO staff played the role of facilitator giving feedback, documentation, materials, and examples until the valid AGRIS AP format was produced. A need for XSLT transformation before inclusion of AGRIS AP to the AGRIS database was identified. It was created and runs at present on each extracted batch file translating ISSN from 9 digits to 2

to fit the specific identifiers of AGRIS (12 characters). This transformation was sent to the SciELO Unit to be integrated in the on-line OAI-PMH plug-in interface and automatically applied to the output.

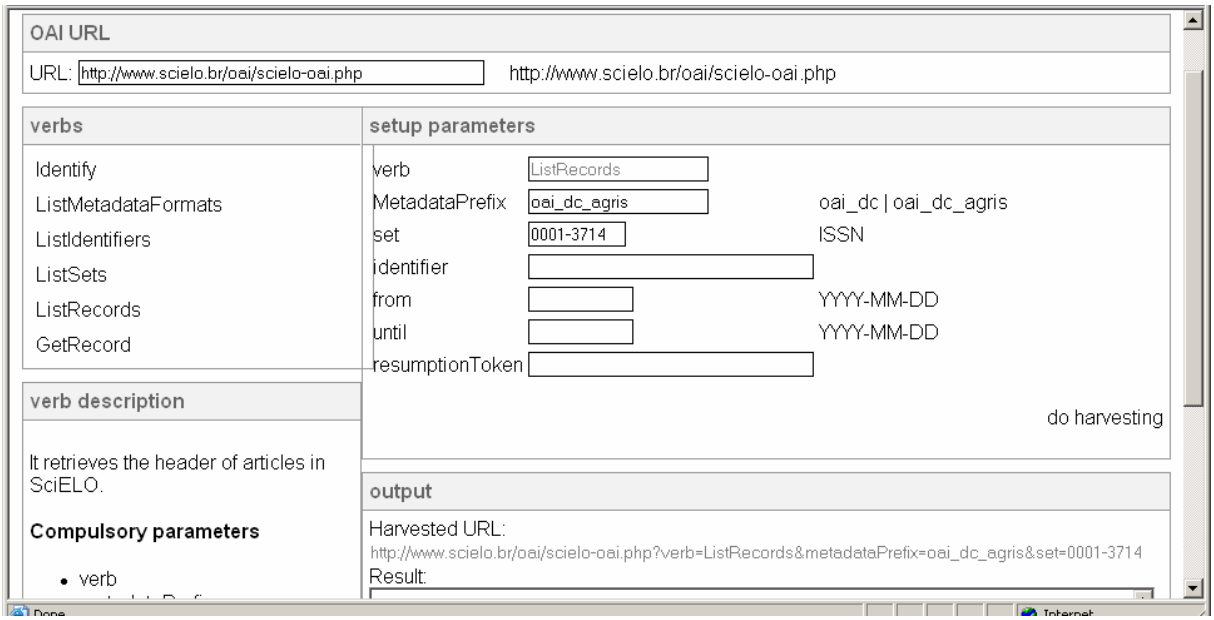
An additional XSLT transformation was required to normalize the OAI-PMH part and produce a valid XML AGRIS AP file, as for example, to include the missing <!DOCTYPE ags:resources SYSTEM "http://purl.org/agmes/agrisap/dtd/"> as well as the namespace definition for the resources.

```
<ags:resources xmlns:ags="http://purl.org/agmes/1.1/" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:agls="http://www.naa.gov.au/recordkeeping/gov_online/agls/1.2"
xmlns:dcterms="http://purl.org/dc/terms/"
xsi:schemaLocation="http://www.purl.org/agmes/agrisap/schema/
http://www.purl.org/agmes/agrisap/schema/agris_ap.xsd">
```

This XSLT was used during manual harvesting at the described stage. It will be used until the AGRIS harvester includes and runs this procedure automatically.
The received AGRIS AP XML files were checked and validated using XMLSpy editor.

3. Testing and implementation of harvesting at FAO

Harvesting great volumes of (legacy) data over the Internet is not an easy process.
The initial process included harvesting legacy data (all selected articles) in AGRIS AP XML files and loading them in AGRIS like it was done for Pubmed. For each journal a batch was produced. For some big files we divided the output in more than one file. After the initial loading of legacy data was done incremental harvesting using OAI-PMH was tested at FAO. It was done in an automatic way using OAI-PMH verb ListRecords. AGRIS AP format was introduced as a parameter in the verbs of OAI-PMH. We applied a script for ListRecords for harvesting SciELO journal articles using different sets (setSpec) identified by the ISSN of the journal in AGRIS AP XML format.



Here are samples of the 6 verbs used for OAI-PMH:

Identify

http://www.scielo.br/oai/scielo-oai.php?verb=Identify

ListMetadataFormats

http://www.scielo.br/oai/scielo-oai.php?verb=ListMetadataFormats

ListSets

http://www.scielo.br/oai/scielo-oai.php?verb=ListSets

ListIdentifiers

http://www.scielo.br/oai/scielo-oai.php?verb=ListIdentifiers&metadataPrefix=oai_dc_agris&from=1997-01-01&until=1999-01-01&set=0001-3714

http://www.scielo.br/oai/scielo-oai.php?verb=ListRecords&metadataPrefix=oai_dc_agris&from=1998-01-01&until=1999-01-01&set=0001-3714

The limit for harvesting at one time was set at 100 records and then Resumption Token was used for the rest of the result. In order to run in batch mode a script including the Resumption Token when the result is divided into more than one batch a parameter &resumptionToken was used:

Example:

http://www.scielo.br/oai/scielo-oai.php?verb=ListRecords&metadataPrefix=oai_dc_agris&from=1999-01-01&until=1999-12-31&set=0001-3714&resumptionToken=HR__S0001-37141999000400013:0001-3714:1999-01-01:1999-12-01

ListRecords

http://www.scielo.br/oai/scielo-oai.php?verb=ListRecords&metadataPrefix=oai_dc_agris&from=1997-01-01&until=1999-01-01&set=0001-3714

GetRecord

http://www.scielo.br/oai/scielo-oai.php?verb=GetRecord&metadataPrefix=oai_dc_agris&identifier=oai:scielo:S0001-37141998000300001

In future harvesting will be done by the harvester in an automatic way. We are now testing the inclusion of AGRIS AP as a possible parameter into the existing harvesters (PKP2 and OCLC Harvester2).

4. Structure of the result harvested through OAI-PMH:

```
<?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>2008-06-23T08:22:32Z</responseDate>
  <request verb="ListRecords" metadataPrefix="oai_dc_agris" from="1997-01-01" until="1999-01-
01" set="0001-3714">http://www.scielo.br/oai/scielo-oai.php</request>
- <ListRecords>
- <record>
- <header>
  <identifier>oai:agris.scielo:BE1998000301</identifier>
  <datestamp>1998-09-01</datestamp>
  <setSpec>0001-3714</setSpec>
  </header>
- <metadata>
```

```
- <ags:resources xmlns:ags="http://purl.org/agmes/1.1/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:agls="http://www.naa.gov.au/recordkeeping/gov_online/agls/1.2"
xmlns:dcterms="http://purl.org/dc/terms/">
- <ags:resource ags:ARN="BE1998000301">
- <dc:title xml:lang="en">
- <![CDATA[
MICROBIAL COUNTS OF DARK RED LATOSOL SAMPLES STORED AT DIFFERENT
TEMPERATURES
 ]]>
 </dc:title>
- <dc:creator>
 <ags:creatorPersonal>Vieira, Francisco Cleber Sousa</ags:creatorPersonal>
 <ags:creatorPersonal>Nahas, Ely(Universidade Estadual Paulista)</ags:creatorPersonal>
 </dc:creator>
- <dc:publisher>
 <ags:publisherName>Sociedade Brasileira de Microbiologia</ags:publisherName>
 </dc:publisher>
- <dc:date>
 <dcterms:dateIssued>1998</dcterms:dateIssued>
 </dc:date>
 <dc:subject>bacteria</dc:subject>
 ...
 <ags:availabilityLocation>SCIELO</ags:availabilityLocation>
 <ags:availabilityNumber>10.1590/S0001-37141998000300001</ags:availabilityNumber>
 </agls:availability>
- <ags:citation>
 <ags:citationTitle>Revista de Microbiologia</ags:citationTitle>
 <ags:citationIdentifier scheme="ags:ISSN">0001-3714</ags:citationIdentifier>
 <ags:citationNumber>vol.29 num.3</ags:citationNumber>
 <ags:citationChronology>1998/09</ags:citationChronology>
 </ags:citation>
 </ags:resource>
 </ags:resources>
 </metadata>
 </record>
```


Results and Future Prospects

1. Overall results and implementation:

For the first time SciELO data were harvested in AGRIS AP format. We have harvested metadata from 30 journals until now. The files are valid and three of them were registered and included in the AGRIS database for search.

We have already included in AGRIS more than 17000 SciELO journal articles. The table below presents some of the journals harvested:

Journal name	ISSN	Number of records
Brazilian Journal of Microbiology	1517-8382	700
Anais da Academia Brasileira de Ciências	0001-3765	995
Anais da Sociedade Entomológica do Brasil	0301-8059	371
Arquivo Brasileiro de Medicina Veterinária e Zootecnia	0102-0935	1291
Revista de Microbiologia	0001-3714	98

The articles from the selected SciELO journals can be seen by using the AGRIS search engine at: http://www.fao.org/agris/search/search.do?query=%2Bcenter%3A%28XS%29

At present we run XSLT transformation in batch mode locally after the harvesting is done until BIREME introduces it in the on-line OAI plug-in. Some adjustments to the OAI-PMH format are still required. The inclusion of SciELO as a data provider in the AGRIS Harvester at FAO is under way.

2. Encountered Problems

Among the problems met during the transformation to the AGRIS AP XML were:

a) some records from SciELO have no subject element at all. This is a problem as subject is a mandatory element for AGRIS AP. A solution could be found by automatic assignment of the subject in case of absence (using, for example, keywords from the journal or article title);

b) AGROVOC is not used in SciELO and the AGRIS semantic tools can not be applied to SciELO data if based on AGROVOC;

c) some forbidden characters were found in the text of the AGRIS AP metadata (HTML tags), which had to be cleaned or enclosed in CDATA in order to produce a valid file;

d) language identification is missing (for example, xml:lang attribute) from the citationTitle, the title of the journal or the subject;

For example:

```
<dc:subject>amplification</dc:subject>
```

should be:

```
<dc:subject xml:lang="en">amplification</dc:subject>
<dc:subject" xml:lang="pt">identificação</dc:subject>
```

e) DOI of the related articles was not included in the list of references in the SciELO XML representation:

```
<dc:relation>
        <dcterms:references scheme="ags:DOI">10.1590/S0001-
371419990002000...</dcterms:references>
</dc:relation>
```

Those links can be seen in the full text presentation of the records;

f) AGRIS and SciELO use different standards for data description. For example, authors' name is Sircili, M. in AGRIS and Sircili, Marcelo Palma in SciELO.

3. Conclusion and Future Prospects

The new format of representation in AGRIS AP of SciELO metadata enriches AGRIS collection and gives more access and visibility to the SciELO journals and their full text articles.

This work shows how the resources from distributed sources can be integrated using the common rules and standards (AGRIS AP and OAI-PMH). OAI-PMH can process different format

of the resources defining and following the common (minimum) requirements (after mapping of the local structure to the AGRIS AP elements).
The selection of a subset of SciELO articles using general criteria (defining journals with agricultural contents) and its integration in AGRIS portal searchable within the common repository and portal is an achievement facilitating the users to do common search in AGRIS and to find SciELO records in addition to the ones from AGRIS.
Once SciELO articles are included in the AGRIS repository it will be possible to carry out more precise search by subject, author, etc. as well as to implement semantic tools created for the AGRIS portal.
Identifying the source of the indexing terms (thesaurus) for SciELO records will give additional possibilities for search query expansion by browsing a vocabulary during the search application.
BIREME's OAI-PMH plug-in can be used (on-line) by any agricultural centre that collects input in AGRIS AP in order to import SciELO records and include them into local databases and local search engines without need to catalogue them again. The experiment with SciELO Brazil can be further expanded by harvesting also other SciELO data providers such as Cuba, Mexico, Chile and Spain

Our experience with this new way of improving visibility and accessibility to SciELO data through AGRIS service providers shows that this important first step towards open access publishing and exchange of common technologies (in this case, between FAO and BIREME) can be used for all other areas of work that are as close as agriculture and health are.

Acknowledgements:
We express our gratitude to the whole BIREME staff for their work and continuous support for achieving these results.
In addition to the authors of this article, the FAO team included: Irene Onyancha and Imma Subirats (selection of journals), Maria Folch (harvesting manually, validating and editing the files), Stefano Anibaldi (validating and publishing the AGRIS AP files in AGRIS repository for search).

References:

[1]Imma Subirats, Irene Onyancha, Gauri Salokhe, Johannes Keizer: Towards an architecture for open archive networks in Agricultural Science and Technology ftp://ftp.fao.org/docrep/fao/009/ah766e00.pdf
[2] Kaloyanova, Stefka, Betti, G., Castellani, F., Keizer, J.. (2008) "Achieving OAI-PMH compliancy fro CDS/ISIS databases" (2008). The Electronic Library, vol. 26, No. 3, ftp://ftp.fao.org/docrep/fao/010/ai156e/ai156e00.pdf
[3]AGRIS AP: The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description (2005). Retrieved March 2007, from http://www.fao.org/docrep/008/ae909e/ae909e00.htm
[4] Salokhe, Gauri, . (2007) "Benefits of AGRIS AP over simple DC in OAI environment" Retrieved March 2007, http://agriscontent.wordpress.com/2007/01/09/benefits-of-agris-ap-over-simple-dc-in- OAI-environment/