



Achieving OAI PMH compliancy for CDS/ISIS databases

Achieving OAI
PMH compliancy

Stefka Kaloyanova

*Knowledge Exchange & Capacity Building Division,
Food and Agriculture Organization of the United Nations, Rome, Italy*

Gian Luigi Betti and Francesco Castellani

*Associazione per la documentazione le biblioteche e gli archivi (DBA),
Florence, Italy, and*

Johannes Keizer

*Knowledge Exchange & Capacity Building Division,
Food and Agriculture Organization of the United Nations, Rome, Italy*

411

Received 12 June 2007
Revised 14 November 2007
Accepted 28 November 2007

Abstract

Purpose – The main purpose of this paper is to present the work recently carried out by the Food and Agriculture Organization of the United Nations (FAO) in collaboration with Associazione per la documentazione le biblioteche e gli archivi (DBA) in Italy to make web CDS/ISIS-based applications compliant with the OAI-PMH. CDS/ISIS is an Integrated Storage and Information Retrieval System of Unesco, which is widely used especially in Latin America and Africa. There are hundreds of CDS/ISIS-based application systems managing bibliographical reference, ensuring high quality content through the use of built-in authority files, data entry guidelines and validations. It also allows for metadata export in many different formats.

Design/methodology/approach – The methodology adopted included study, analysis and evaluation of three existing solutions for exposing metadata from the CDS/ISIS database repositories to the OAI framework.

Findings – The implementation did not include the development of automatic procedures for incremental harvesting from CDS/ISIS databases nor the normalization of the harvested data. However, a lot of experience in implementation of OAI was gained which will be useful for future development of non-CDS/ISIS systems.

Research limitations/implications – The research and development work demonstrates the importance and implications of this work for the whole CDS/ISIS community and specifically for the participating centres from the AGRIS network.

Originality/value – It proposes an open source, easily parametrizable plug-in tool, which can be adapted to expose metadata from a general structure CDS/ISIS database using the OAI-PMH protocol. This work assures that semantically rich metadata for agricultural science and research publications based on the “AGRIS Application Profile” can be handled by the OAI protocol. This in turn allows for further creation of additional services based on the exchange of knowledge on agricultural science and technology publications world-wide.

Keywords Archives management, Databases, Information retrieval

Paper type Technical paper



© Food and Agriculture Organization of the United Nations, 2008

The authors would like to thank all the team participating in this work and especially: Jeff Joung (OCLC) for OAICat servlet development; Mariusz Suchorowski (ICIE Poland) and Stefano Pesci for their support; Irene Onyancha and Gauri Salokhe for editing this article and for metadata crosswalks and testing; Stefano Anibaldi for metadata crosswalks and testing; Kat Hagedorn (OAIster) for harvesting exposed metadata; and Steve Katz for overall support and guidance.

Introduction

CDS/ISIS is an Integrated Storage and Information retrieval System of United Nations Educational Scientific and Cultural Organization (Unesco), which is widely used throughout the world especially in Latin America and Africa for creating metadata management systems. There are hundreds of CDS/ISIS based application systems managing bibliographical reference ensuring high quality content through the use of built-in authority files, data entry guidelines and validations (CDS/ISIS Computerized Documentation Systems/ Integrated Set of Information Systems http://portal.unesco.org/ci/en/ev.php-URL_ID=2071&URL_DO=DO_TOPIC&URL_SECTION=201.html). The web based CDS/ISIS applications (WWWISIS/wxis (products of BIREME) and WWW-ISIS[1] based interfaces) evolved from library catalogues to interoperable systems, operating on metadata and associated digital objects. The features of these two systems guarantee high quality metadata content based on validation tools, built-in authority files as well as compliancy to common standards which include Dublin Core (DC) (<http://dublincore.org/>), AGRIS Application Profile (AGRIS AP) (www.fao.org/agris/tools/AGRIS_AP/WhatItIs.htm), AGROVOC [2], AGRIS Subject categories (www.fao.org/Agris/) etc. These applications are appreciated as an easy to use solution towards achieving high quality metadata with links to digital full text documents. They can be easily adapted to new platforms, systems and standards using flexible integrated formatting features (Kaloyanova and Okoniewski, 2005) and produces DC or AGRIS AP compliant output in many formats including: Extensible Markup Language(XML)(www.w3.org/XML/), Hypertext Markup Language (HTML) (www.w3.org/TR/xhtml1/), Resource Description Framework (RDF) (www.w3.org/RDF/),tag/delimited etc.

CDS/ISIS is also used within FAO for the FAO on-line catalogue (FAOBIB) (www4.fao.org/faobib/). The WebAGRIS [3] application, based on WWW-ISIS, is one used by many centres within the AGRIS network (Rybinski *et al.*, 2005). As shown by the statistics from the total 54 active AGRIS network participating centres, 50 centers are using CDS/ISIS based applications. WWWISIS/wxis of BIREME (www3.bireme.br/bvs/bireme/E/homepage.htm) is another CDS/ISIS application used by many institutions in Latin America and the Caribbean forming part of SIDALC (www.sidalc.net). The general observation is that the users are comfortable with these applications and do not want or do not have the resources to move to new systems.

The need of CDS/ISIS users to achieve Open Archives Initiative (OAI) interoperability without the obligation to change the existing was the main motivation of this development. The objective of this work was to achieve full OAI-PMH (OAI-PMH) (www.openarchives.org/pmh/) compliancy for systems based on the CDS/ISIS software. This would have to be introduced through an open source solution, which in turn has to be a software layer (plug in) that permits the harvester to pull data from Internet accessible CDS/ISIS databases using OAI-PMH with different metadata schemas.

Methodology

The methodology that we adopt included study, analysis and evaluation of some of the existing solutions for exposing of metadata from the CDS/ISIS database repositories to the OAI framework. We evaluated the three existing studies as follows:

- (1) The solution described in the paper "A Dynamic approach to make CDS/ISIS database interoperable over internet using OAI protocol" (Jayakanth *et al.*,

2006). Following the proposed solution some trials with FAO metadata were done, and the results indicate that it is a dynamic solution but only for databases that conform to machine readable cataloguing (MARC) standard and not a general structure CDS/ISIS applications. Any other non-MARC format structured CDS/ISIS database had to be converted first to an intermediary (model) database, using reformatting field select table (FST) for changing the structure of the original application.

- (2) A PHP-based tool created at BIREME within the “SciELO” system (www.scielo.org/index.php?lang=en) also revealed that it was not a general purpose application that was easy to adapt to different CDS/ISIS database structure.
- (3) The experience of DBA (www.dba.it), managing wxis CDS/ISIS databases, which was based on the extension of OAICat of Online Computer Library Center (OCLC) (an open source Java Servlet Web application for exposing databases to OAI-PMH v2.0 framework). OAICat was adapted to run with Web CDS/ISIS wxis script (BIREME) as an intermediary between a data provider (CDS/ISIS database) and an OAI harvester.

The conclusion was that the first two solutions were not suitable for the general structure CDS/ISIS database applications neither for the AGRIS centres that want to remain with WEBAGRIS and still be compliant to OAI. We needed an interface that could be adapted to any structure of CDS/ISIS database by changing only a few parameters and not reformatting continuously to MARC format structured database.

The third solution mentioned above met the set requirements and was chosen to be the base for development of a technology of exposing the CDS/ISIS repositories in the AGRIS Open Archive Networks. It was developed further to be easily parameterized and be a more general purpose tool based on the OAICat for dynamic harvesting of metadata from the CDS/ISIS Web applications. It had to be also adapted to:

- general WWW-ISIS-based applications (for example, WEBAGRIS applications in AGRIS network); and
- WWWISIS/wxis of BIREME based applications used by DBA and some participating in SIDALC centres.

The experience accumulated from distributed search over heterogeneous repositories indicates that it is simpler to aggregate the metadata and then search it because the storage becomes cheaper. The Open Archives Initiative[4] provides a way of building such aggregation using harvesters based on the Protocol for Metadata Harvesting (OAI-PMH). The OAI framework and the technical specifications for its metadata harvesting were the base for this development. OAI framework participants are data providers and service providers. The OAI-PMH principle is to transport from open accessible distributed data providers common syntax XML- encoded byte stream that serves as a packaging mechanism for harvested metadata to an XML repository or to a federated search at server provider site. The content normalization and the extended usage of the harvested metadata is more at the service provider site. A repository is a network accessible server that can process the 6 OAI-PMH verb requests and managed by a data provider that exposes metadata to harvesters. A harvester is operated by a service provider as a means of collecting metadata from repositories by issuing OAI-PMH requests.

Harvesting process flow

The OAI harvester (at the service provider) sends an OAI protocol request (query) to the servlet program. The program parses the query and parameters and translates the query to a script meaningful for WWW-ISIS or wxis application. Then the script is executed by the local application (WWW-ISIS or wxis) and after interaction with the database the matched records are formatted according to the required specification and passed back by the servlet program to the harvester in appropriate XML file that follows the syntax of unqualified DC metadata format (OAI_dc scheme) or AGRIS AP XML (agris_ap scheme) for record presentation. In this process the local application has to ensure execution of the queries in the form of scripts, including in advance the search criteria in the inverted file of the CDS/ISIS database as well as formatting of the result records, using the CDS/ISIS formatting language as required. The process flow is indicated in Figure 1. The CDS/ISIS application should ensure mandatory elements (datestamp, sets) for its selective harvesting in an automatic incremental mode (new updates).

The quality of service is proportional to the quality of the data harvested. Assuring a good quality metadata requires that appropriate data management tools are used. Web based CDS/ISIS systems that ensures such a quality data collection as well as compliancy of the Institutional Repositories to OAI was a good starting point.

Description of work done

In addition to the building institutional repositories (using CDS/ISIS based applications) building of other components of the OAI framework was done at the data provider site. In more details the following steps were taken:

- (1) Development of plug-in.
- (2) Adaptation of the existing CDS/ISIS application (field select table, metadata formats for formatting and sorting the data elements of the response in appropriate XML file) according to the OAI request records.
- (3) Ensuring XML schema validations (namespace definition): supporting DC simple and AGRIS AP.
- (4) Preparation and testing of search formulation (script) equivalent to the harvester request, ensuring all required parameters of OAI-PMH.
- (5) Registration of the repository and supplying of URL for the repository at the OAI framework (repository and record identification).
- (6) Installation of ISISOAI servlet.

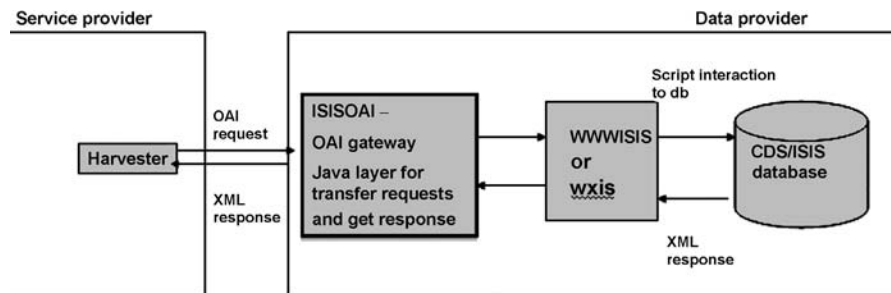


Figure 1.
Flow of the process

These steps are described in more detail below.

As a first step, a software layer (ISISOAI plug-in) was created as mediator between a CDS/ISIS database (at data provider) accessible on the internet and a harvester (at the service provider site) to ensure exposing of CDS/ISIS databases (FAOBIB and WEBAGRIS) to harvesters according to OAI-PMH protocol requirements. ISISOAI is an Open source Java Servlet web application to query and expose a generic CDS/ISIS database to OAI-PMH v2.0 framework. It is an extension of the OAICat application (<http://pubserv.oclc.org/oaicat/jars/docs/index.html>) and runs under an application server compliant Java Platform, Enterprise Edition (J2EE) (<http://java.sun.com/javaee/>) (e.g. Apache's Tomcat). It is adapted for use with WWW-ISIS and wxis applications and can further be adapted for accessing repositories stored as an XML file or in a relational database management system (RDBMS).

After mapping the local metadata scheme to common AGRIS AP and DC exchange formats CDS/ISIS print formats for record presentation were prepared at the CDS/ISIS part. Then, scripts equivalent to the harvester's requests were created in wxis or WWW-ISIS and tested including parameters for the script, which reflects the harvester request.

The structure of an XML record exposed from the database to a harvester is comprised of two mandatory parts:

- (1) *Header* – (independent of the metadata format of the record) representing the common part necessary for the harvesting process, including unique identifier (key) for extracting metadata from a specific record in a specific repository (oai:fao.agris:UY20071000001), datestamp (date of the last access to the record) and sets (specification of subsets by type of records or subject).
- (2) *Metadata* – the OAI technical framework mandates the use of simple DC metadata. Simple DC syntax has much larger requirements. All its elements could be repeatable and optional. Most of the OAI services are based on this format. The original data elements are grouped under 15 main DC elements. Simple DC structure is not meant for storage but to be used for dynamic conversion to and as exchange format in response to a harvester request. Using DC in the OAI community is the first step towards OAI based IR (Institutional Repositories) interoperability for more global acceptance. However, this is not sufficient for the Agricultural community. That is why it was important to create and implement a specific metadata set with the corresponding schema AGRIS AP in order to improve the quality of services within the agricultural community.

The AGRIS AP syntax[5] 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, 2007). The value of the extended common metadata standards is achieved at the service provider level where many value added services are possible. However, at the data provider site it may create more difficulties in validating, especially legacy data.

An optional part "about" will be also included for data about the metadata such as provenance of the records. This is very important in a network community for identifying the provenance of the records so that its uniqueness in the OAI framework can be guaranteed.

The next step was the installation of ISISOAI servlet and testing exposure of metadata in response to a request, This included executing of the six verbs of the OAI Metadata Harvesting Protocol namely: “Identify” to retrieve information about repository, “GetRecord” which is used to retrieve an individual record by unique identifier in the required format, “List identifiers” to retrieve the identifiers of records that can be harvested (this could include a subset selection, “ListMetadataFormats” (showing metadata formats available for the repository), “ListRecords” (to harvest records from a repository in appropriate format) and “List Sets” (to retrieve the set structure in a repository). ListSets is a method of exposing a part (subset) of the repositories contents to harvesters. It depends on the harvester to exploit this capability as the implementation of sets differs for different implementers which could be applied in particular community specific cases (for example semantic subsets) (see Figures 2 and 3).

The work included development of a broad range of possibilities for selective harvesting based on specifications for reasonable subset creation based on:

- (1) *Date* – for selection records input/updated within a range of dates (for example, to extract only the new and lately upgraded records since the previous harvesting; used for incremental harvesting as well as for keeping upgraded version of the metadata of a given record); or
- (2) *Sets* – for selection of records by subject, type etc. and selecting subsets (groups) from the selective harvested data. Example: for FAOBIB repository we defined four possible subsets for extraction of:
 - all FAO documents;
 - only records for AGRIS;
 - library books; or
 - full text available FAO documents.



FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

[Identify](#) | [GetRecord](#) | [ListIdentifiers \(Resumption\)](#) | [ListMetadataFormats](#) | [ListRecords \(Resumption\)](#) | [ListSets \(Resumption\)](#)

OAI ListRecords Request Form

from:	<input type="text"/>	From date in the format YYYY-MM-DD
until:	<input type="text"/>	Until date in the format YYYY-MM-DD
set:	<input type="text"/>	Sets defined: X-FAO documents; D-Full text FAO documents; A-AGRIS documents
metadataPrefix:	<input type="text"/>	MetadataPrefix: oai_dc or agris_ap
<input type="button" value="Submit Query"/>		



Jeff Young
Last modified: Tue Oct 14 10:27:15 EDT 2003

Figure 2.
Testing of the verb
ListRecords


```

<?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>2007-05-23T15:52:23Z</responseDate>
  <request identifier="oai:agris.fao:XF2006424587" metadataPrefix="agris_ap"
    verb="GetRecord">http://www4.fao.org:8080/oaiat/OAIHandler</request>
- <GetRecord>
- <record>
- <header>
  <identifier>oai:agris.fao:XF2006424587</identifier>
  <timestamp>2006-01-17</timestamp>
  <setSpec>X</setSpec>
  <setSpec>A</setSpec>
  <setSpec>D</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/"
  xsi:schemaLocation="http://www.purl.org/agmes/agrisap/schema/
    http://www.purl.org/agmes/agrisap/schema/agris_ap.xsd">
- <ags:resource ags:ARN="XF2006424587">
  <dc:title xml:lang="en">Information requirements on hides and skins</dc:title>
- <dc:creator>
  <ags:creatorCorporate>FAO, Rome (Italy). Committee on Commodity Problems</ags:creatorCorporate>
  <ags:creatorCorporate>ESC</ags:creatorCorporate>
  <ags:creatorConference>FAO Committee on Commodity Problems. Intergovernmental Group on Meat.
    Sub-Group on Hides and Skins, Sess. 9, Arusha (United Republic of Tanzania), 1-3 Feb
    2006</ags:creatorConference>
  </dc:creator>
- <dc:publisher>
  <ags:publisherPlace>Rome (Italy)</ags:publisherPlace>
  </dc:publisher>
- <dc:date>
  <dcterms:dateIssued>Nov 2005</dcterms:dateIssued>

```

Figure 3.
Result string for AGRIS
AP ListRecords

Further specifications for subsets by type of document or semantic selection using controlled vocabularies, thesauri or ontology are under development. This will be very useful particularly for very large repositories like FAO's in order to avoid overloading during harvesting of the whole repository but a subject subset of it.

The final steps were validation and registration of the data provider at OAI and other registries. FAO CDS/ISIS metadata from FAOBIB was tested for exposing FAO full text documents which has associated full text documents using simple DC metadata to external OAI harvesters (www4.fao.org:8080/oaiat). Metadata harvesting and the services of the harvested data was carried by OAIster (www.oaister.org/). This was followed by searching over harvested data.

The implementation did not include the development of automatic procedures for incremental harvesting from CDS/ISIS databases, the normalization of the harvested data and creation of repository and common index for federated searching at the harvester (service provider), and the extension to other formats apart from the Simple DC metadata and AGRIS AP XML formats used.

Results and impact

The main result is ensuring an ISISOAI servlet (plug-in) that can be easily installed and adapted for any internet accessible data provider within the AGRIS network that

uses WEBAGRIS or any other CDS/ISIS web application in order to expose metadata to the harvesters.

For the first time general structured CDS/ISIS web applications were open for dynamic harvesting to the OAI framework. In addition, OAI functionality was applied to both simple DC and AGRIS AP syntax which will also be a prerequisite for the development of many new services over high-quality metadata. The experiment with exposing FAODOC metadata through OAIster once again shows the importance of the usage of qualified metadata schema for higher quality of the services.

The ISISOAI was tested with different CDS/ISIS applications:

(1) *WWW-ISIS*:

- FAOBIB the FAO documentation catalogue, an instance of the WWW/ISIS software and exposed through OAIster for search full text FAO documents, exposing about 170,000 records in four sets; and
- WEBAGRIS applications exposing metadata for harvesting to the central AGRIS repository or to other harvesters and service providers.

(2) *WXIS*:

- DbA (sbagnet (Biblioteca dell' Anglona Gallura), Scire, TECAWEB etc.); and
- SIDALC members using CDS/ISIS applications (on going).

In general, a lot of experience in implementation of OAI was gained. This will be useful for future development of non-CDS/ISIS systems.

The impact of this experiment is enormous. The implementation of the results will considerably improve the visibility of FAO and AGRIS network documents and improve the services and functionalities on the Service providers' sites. In addition, with ISISOAI, the whole harvesting process to the AGRIS central XML repository will be reorganized by cutting E-mails and overloads. A harvester for collecting and upgrading AGRIS AP metadata in an XML repository accessible over the internet can be installed at FAO HQ. It can then aggregate metadata in AGRIS_AP from all CDS/ISIS data providers such as Kenya Agricultural Information Network (KAINet)(<http://agriscontent.wordpress.com/2007/03/15/kenya-agris-pilot-project-kenya-agricultural-information-network-kainet/>), AGRORED Peru (www.agroredperu.org/), and ORTON IICA/CATIE Library (<http://orton.catie.ac.cr/bco/>) (see Figure 4).

Further, the extension of the metadata interoperability and search based on the AGRIS AP syntax together with the use of thesauri and ontologies for expressing semantic relations for value-added multilingual searches will be more realistic. And finally, with the same servlet any service provider can extend the use of Simple DC XML format metadata beyond the AGRIS community by including it in a common search interface.

Next steps

Some of the next steps are:

- (1) Preparation of a distribution and installation documentation package.
- (2) Installation of the plug-in to data providers and testing the exposure of metadata through OAI-PMH.

Notes

1. WWW-ISIS software developed by ICIE and CC with strong support from FAO WAICENT (both financial and conceptual); website: <http://w2isis.icml9.org/activity.php?lang=en&id=33>
2. AGROVOC The multilingual structured thesaurus of all subject fields in Agriculture, Forestry, Fisheries, Food security and related domains by the Food and Agriculture Organization of the United Nations AGROVOC website: www.fao.org/aims
3. WEBAGRIS The complete, multilingual Web-based system for distributed data input, processing and dissemination of agricultural bibliographic information website: www.fao.org/agris/tools/WebAGRIS/WebAGRIS_En.htm
4. See Carl Lagoze, Herbert Van de Sompel, "The Open Archives Initiative: Building a low-barrier interoperability framework", available from www.openarchives.org/documents/jcdl2001-oai.pdf
5. 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 from www.fao.org/docrep/008/ae909e/ae909e00.htm

References

- Jayakanth, F., Maly, K., Zubair, M. and Aswath, L. (2006), "A dynamic approach to make CDS/ISIS database interoperable over internet using OAI protocol", available at: <http://eprints.iisc.ernet.in/archive/00008252/01/cdsdynamic04apr06.pdf>
- Kaloyanova, S. and Okoniewski, M. (2005), "FAO's experience in metadata exchange from CDS/ISIS bibliographic databases using XML format, compliant to Dublin Core standard", ICML (2005), available at: <http://w2isis.icml9.org/activity.php?lang=en&id=34>
- Rybinski, H., Kaloyanova, S. and Katz, S. (2005), "WWW-ISIS: a result of a close cooperation between FAO-GIL and ICIE", ICML (2005), available at: <http://w2isis.icml9.org/activity.php?lang=en&id=33>
- Salokhe, G. (2007), "Benefits of AGRIS AP over simple DC in OAI environment", available at: <http://agriscontent.wordpress.com/2007/01/09/benefits-of-agris-ap-over-simple-dc-in-OAI-environment/> (accessed March 2007).

Further reading

- Marcondes, C. and Sayão, L. (2003), "The SciELO Brazilian Scientific Journal Gateway and Open Archives. A report on the development of the SciELO-Open Archives data provider server", *D-Lib Magazine*, Vol. 9 No. 3, available at: www.dlib.org/dlib/march03/marcondes/03marcondes.html

About the authors

Stefka Kaloyanova currently works as an Information Systems Analyst in the Knowledge Exchange and Capacity Building Division at the Food and Agriculture Organization in Rome where she has worked in other capacities since 1991. Prior to that she was Information Systems Manager at the Central Institute for Scientific and Technical Information (CISTI) in Sofia, Bulgaria. She graduated in Mathematics from Sofia University and holds a postgraduate diploma in Librarianship and Studies in Library Automation and Computer Programming. During her (more than 30 years) working experience in information management her main duties and responsibilities have included: building international network environments for data exchange and retrieval, promoting and implementation of common standards, Integrated Library Management Systems development and implementation etc. More recently her main

work involves building Institutional Repositories using an OAI framework and Open Access publishing. Ms Kaloyanova has published and presented to the conferences and congresses many articles on Integrated Library Management Systems, CDS/ISIS applications, Open Archive Initiatives and OAI-PMH implementation. She also acts as referee to several international journals in these areas, including *The Electronic Library* and *ComSIS Journal*. She is the corresponding author and can be contacted at: stefka.kaloyanova@fao.org

Gian Luigi Betti is Information Manager at the Associazione per la documentazione le biblioteche e gli archivi (DBA), Florence, Italy.

Francesco Castellani is an Engineer and Technical Manager at the Associazione per la documentazione le biblioteche e gli archivi (DBA), Florence, Italy.

Johannes Keizer has PhD in Biology from the University of Mainz (Germany) and is working as an Information Systems Officer at the Food and Agriculture Organization of the UN where he is responsible for scientific documentation systems and knowledge and information management standards.