The FAO Open Archive: Enhancing Access to FAO Publications Using International Standards and Exchange Protocols

Claudia Nicolai; Imma Subirats; Stephen Katz

Food and Agriculture Organization of the United Nations
Viale delle Terme di Caracalla 1, 00153 Rome, Italy
e-mail: Claudia.Nicolai@fao.org; Imma.Subirats@fao.org; Stephen.Katz@fao.org

Abstract

Since 1998, the Food and Agriculture Organization of the United Nations (FAO) has been publishing its electronic publications in the FAO Corporate Document Repository (CDR). The electronic publishing workflow is maintained by the Electronic Information Management System (EIMS). The EIMS-CDR holds more than 38 500 documents and is the gateway to FAO's publications. The EIMS-CDR coexists with the FAODOC – the online catalogue for documents produced by FAO. FAODOC catalogues and indexes both electronic and printed documents while the EIMS-CDR manages full text documents and a minimal set of metadata. This paper discusses the merger of the EIMS-CDR and the FAODOC into a unique FAO Open Archive based on the integration of the electronic publishing and the bibliographic cataloguing requirements. The FAO Open Archive will be the foundation for the collection, management, maintenance and timely dissemination of material published by FAO. To improve the effectiveness of the proposed repository, it is necessary to streamline the current electronic publishing workflow. The merger of the EIMS-CDR and the FAODOC will strengthen FAO's role as a knowledge dissemination organization. Especially, as one of the principal tasks of the FAO is to efficiently collect and disseminate information regarding food, nutrition, agriculture, fisheries and forestry.

Keywords: open access; open archive initiative; interoperability; digital repositories; data content standards

1 Introduction

The Food and Agriculture Organization of the United Nations (FAO) has more than 50 years of experience in the production and the dissemination of information, both through its headquarters-based regular programme and through field projects. The collection, analysis, interpretation and dissemination of information relating to nutrition, food and agriculture are FAO's main functions [1]. The World Wide Web has proven to be a powerful means for FAO to disseminate multilingual information.

In this context, FAO was an early implementer of:

1. an online catalogue for documents produced by FAO (FAODOC, Figure 1), a multilingual online catalogue which contains bibliographic metadata of FAO electronic and printed documents [2];
2. the Electronic Information Management System (EIMS), a workflow management tool and database which manages the publication of electronic documents and multimedia resources on FAO's Web sites [3]; and
3. the Corporate Document Repository (CDR, Figure 2), a corporate output interface for FAO full text electronic publications stored in the EIMS [4, 5].

The FAODOC is a multilingual, online catalogue of documents and publications produced by FAO since 1945. The system uses UNESCO's CDS/ISIS software [6]. More than 160 000 documents have currently been catalogued. Since its inception, the FAODOC has focused on the production of high quality bibliographic records.

The FAO Web site was released in 1995 and the first electronic publishing workflow (through EIMS) was initiated in 1998. Currently, more than 38 550 resources (full text documents and multimedia items) are managed by the EIMS (Table 1). Photos, videos and audio are accessible through different systems on the FAO Web site. The CDR was conceived as the online digital library of FAO electronic documents and publications, as well as selected non-FAO material. At present, more than 23 000 full text documents are available through the CDR.

Resource type	Number of Records
full text documents	23 000
photos	8 500
videos	6 300
audio	750
Total	38 550

Table 1: Resources at FAO (as at 10 April 2007)

For each system described above, the objectives are different. The FAODOC focuses on the cataloguing of FAO documents. The EIMS deals with electronic publishing, especially the management at the full text level (rather than the description of documents). The CDR focuses on the dissemination of FAO documents archived through the EIMS. In 2003, a link between both databases was created, linking the FAODOC records to the full text documents archived in EIMS-CDR.

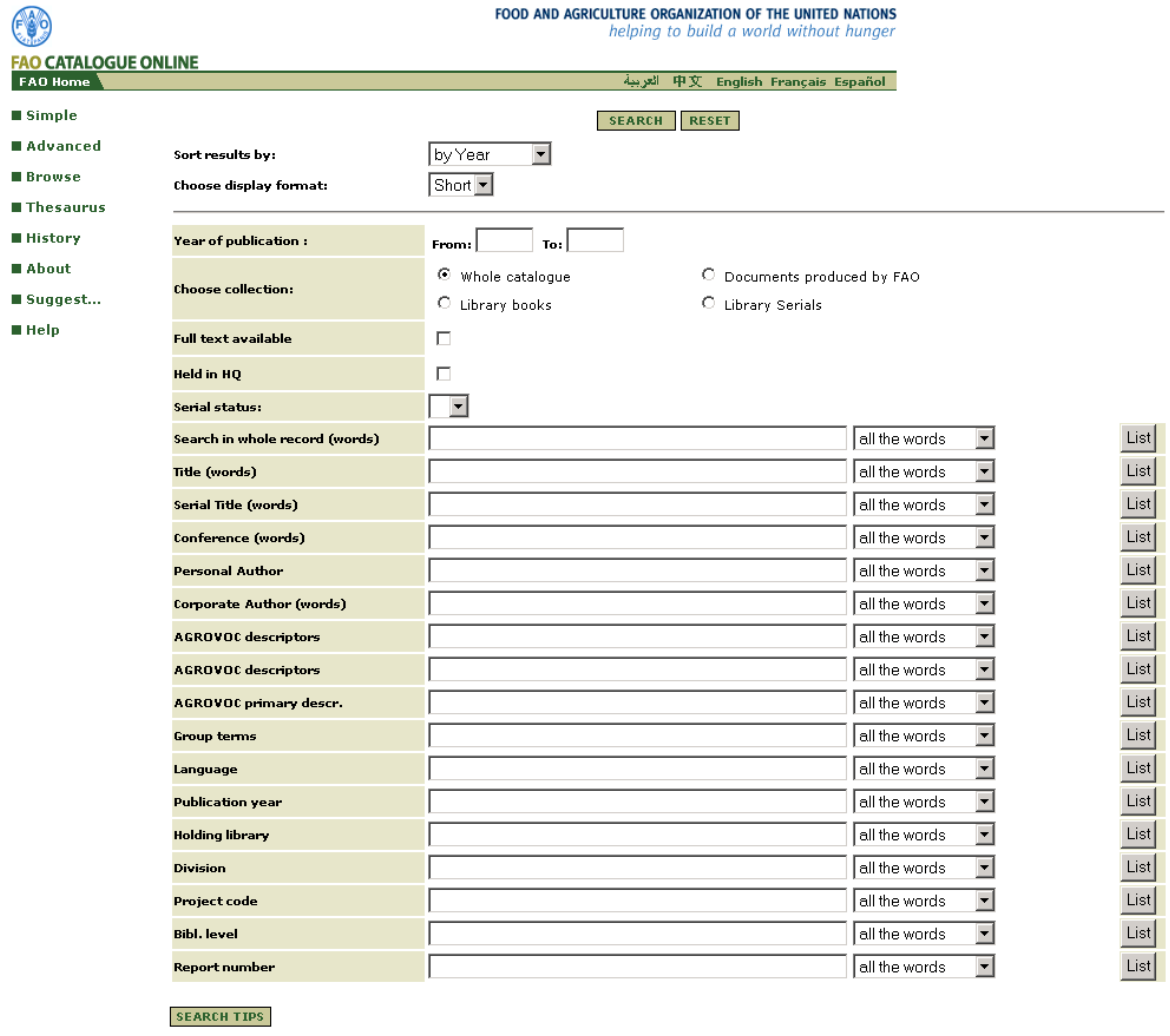

Figure 1: FAODOC user interface

This paper describes the process of merging the EIMS-CDR and the FAODOC and the creation of the FAO Open Archive. The result will be one unique sustainable digital repository offering a solid foundation for the collection, management, maintenance and timely dissemination of material published by FAO. To improve the effectiveness of the proposed repository, it will be necessary to streamline the existing electronic publishing

workflow and to integrate the current functions into new modules. The FAO Open Archive is based on three key elements:

1. a metadata set based on international description guidelines and format;
2. a workflow procedure that guarantees the processing of all documents published by FAO; and
3. a system architecture based on cataloguing and electronic publishing.

This paper is divided into the following sections: Section 2 presents the current situation for the EIMS-CDR and the FAODOC; Section 3 details the objectives of the FAO Open Archive; Section 4 describes the workflow procedures, the new architecture, the compliance to International Standards for Bibliographical Description (ISBD) [7] and metadata sharing with other systems; and Section 5 is the conclusion and the next steps in implementing the FAO Open Archive.

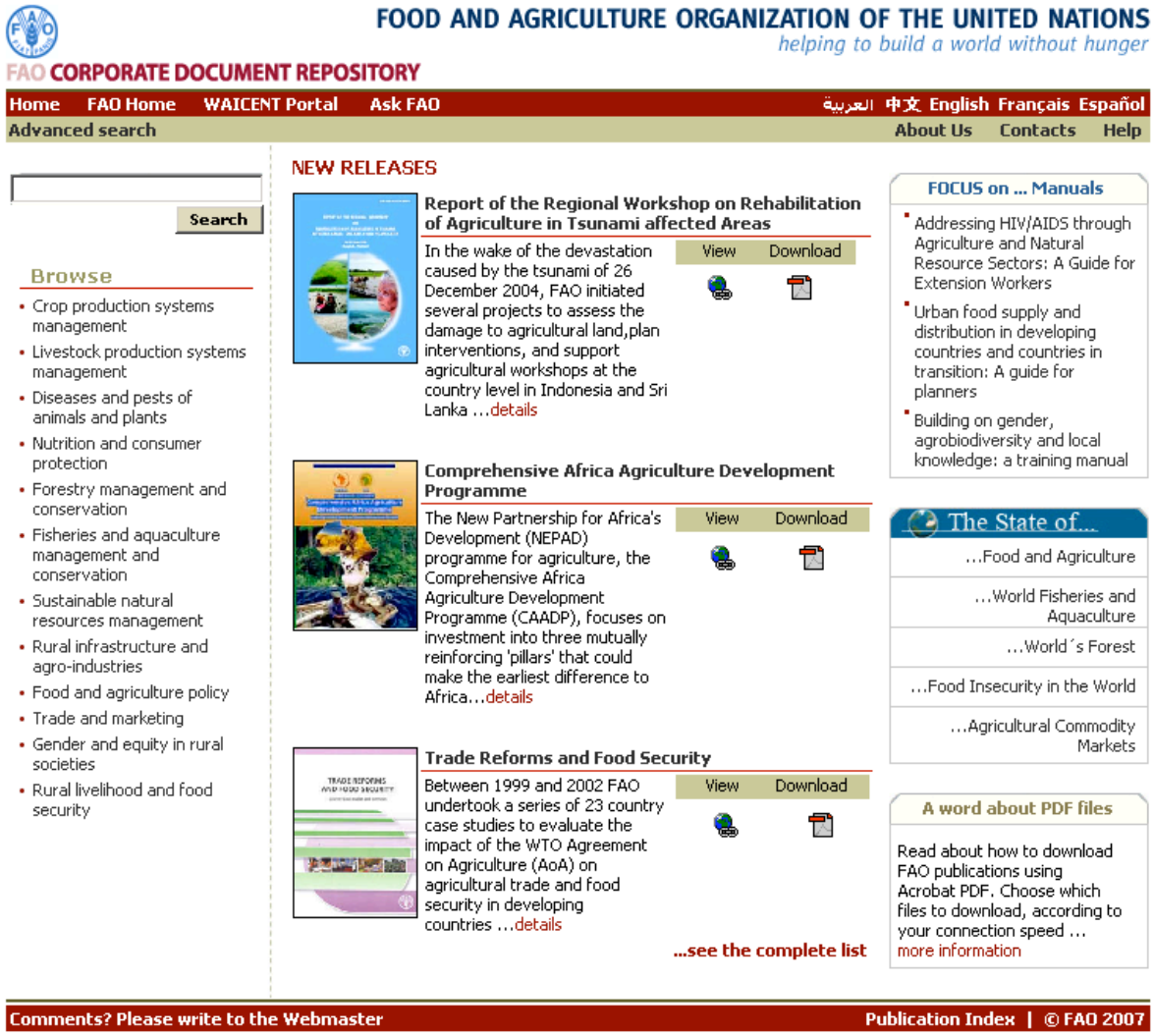

Figure 2: CDR user interface

2 Objectives

The objective of the FAO Open Archive is to create a unique sustainable digital repository for the dissemination of FAO publications and simultaneously, enhance interoperability with other information systems. The FAO Open Archive will guarantee efficient electronic publishing and metadata management, the effective dissemination of FAO information resources and the preservation of the Organization's institutional memory.

3 Current Situation for EIMS-CDR and FAODOC

FAODOC has been managing all bibliographic information for FAO documents and publications for over 30 years (since 1976). Since 1998, FAO established a workflow to manage the electronic publishing and dissemination of FAO full text documents through the EIMS-CDR [8]. The EIMS-CDR and the FAODOC workflows, actors and content are described below.

3.1 EIMS-CDR, the Electronic Publishing and Digital Repository

There are four different user profiles in the EIMS-CDR workflow:

- originator – the person within the FAO unit responsible for providing the source files and/or the printed copy of the publication;
- data owner – the FAO unit responsible for the content of the publication;
- focal point – the person responsible in EIMS-CDR for managing requests from FAO units [9]; and
- liaison officer – the person within a FAO unit who ensures that publications are made available online. The liaison officer is the link between the originator and the focal point.

Detailed guidelines of the EIMS-CDR workflow are available to all FAO users and EIMS-CDR administrators. Following is a brief description of standard workflow steps:

1. The originator provides source files to the external printing unit. When the publication is printed, the external printing unit provides the focal point with the source files, the PDF version and the hard copy. In some cases files are provided by the originator;
2. The data owner creates and locates a record in EIMS;
3. The data owner notifies the focal point of the record and the uploaded files;
4. The focal point completes the record. Conversion to HTML or PDF is handled by focal points or outsourced to an external company. When conversion is completed, the focal point notifies the data owner of the test URL for reviewing the publication;
5. The data owner reviews the publication and either approves it or requests changes, by notifying the focal point;
6. The focal point reviews the final publication, publishes it and notifies the data owner of the public URL. If no conversion is required, the focal point prepares an HTML table of contents that links to the low-resolution PDF files and notifies the data owner of the public URL (in some cases only PDF files are published without the associated HTML pages).

Publications are made available in various electronic formats:

- Full HTML version; HTML loads quickly and is easier to read on-screen. ~14 000 records;
- Full PDF version; PDF is better for printing and downloading a local copy. ~2 200 records;
- Full HTML version and PDF version. ~6 500 records; and
- HTML table of contents linked to Full PDF version. ~500 records

3.2 FAODOC, the Online Catalogue

The FAODOC cataloguing process involves various actors:

- originator – the person within the FAO unit responsible for delivering to FAODOC the hard copy of the publications and/or the full text documents to be published in EIMS-CDR;
- EIMS-CDR focal point – the person who notifies the FAODOC cataloguer of a new record in EIMS-CDR, so they link the FAODOC record to the EIMS-CDR full text document; and
- cataloguer – the person who selects and catalogues the publications (hard copies and full text documents from EIMS-CDR).

The FAODOC manages the cataloguing of document and the dissemination of bibliographic information through an Online Public Access Catalogue (OPAC). There are procedures for the exchange of information between the FAODOC and the document producers, but there is no specific electronic tool to manage the reception of documents, as exists in the EIMS-CDR workflow. The lack of any workflow management system makes it difficult to control the reception and cataloguing of documents.

3.3 Main Differences between EIMS-CDR and FAODOC

The process of merging the two existing databases is a challenging task, as each has a different structure and workflow procedure. The first step towards the FAO Open Archive was to determine the similarities and differences between the EIMS-CDR and the FAODOC.

3.3.1 Software Overview

The EIMS-CDR was developed by FAO to manage the electronic publishing workflow. The CDR and the EIMS both run on a Microsoft Windows platform with an Oracle 9 database server. The software uses Microsoft's ASP programming language (Active Server Pages), with some ad hoc modules and functionalities developed in ASP.Net (the successor to ASP). The EIMS architecture results from the interaction of several modules that manage different aspects of the overall workflow. All modules interact with a single database that stores the records' descriptive metadata and detailed workflow information.

The FAODOC uses CDS/ISIS, a software package for information storage and retrieval – developed, maintained and disseminated by UNESCO. It is freely available for non-commercial purposes. The customization of data input and output interfaces occurred in Poland at the Institute for Computer and Information Engineering and at FAO.

3.3.2 Metadata Structure

CDS/ISIS manages a database whose main content is text, while the EIMS-CDR uses a relational Oracle database. The structure and logic of the two databases are completely different. However, these differences are not a barrier for the merger into a new single relational database.

Both systems use a very similar set of metadata fields to describe documents. The FAODOC contains detailed document information, while the EIMS-CDR provides fewer details on the actual document, but stores much information related to the actors, workflow and full text management. The mapping of the EIMS-CDR and the FAODOC databases has already occurred. It was not a complicated procedure, as both systems use a similar metadata field set. The compliance of both databases to the Dublin Core metadata standard and the AGRIS AP [10] at export level, facilitated the mapping. Only those fields required for the EIMS-CDR workflow have been added to those that already exist in the FAODOC.

3.3.3 Database Content

The EIMS-CDR and the FAODOC currently use FAO cataloguing guidelines. The decision to adopt international cataloguing standards was taken to guarantee interoperability with other digital repositories.

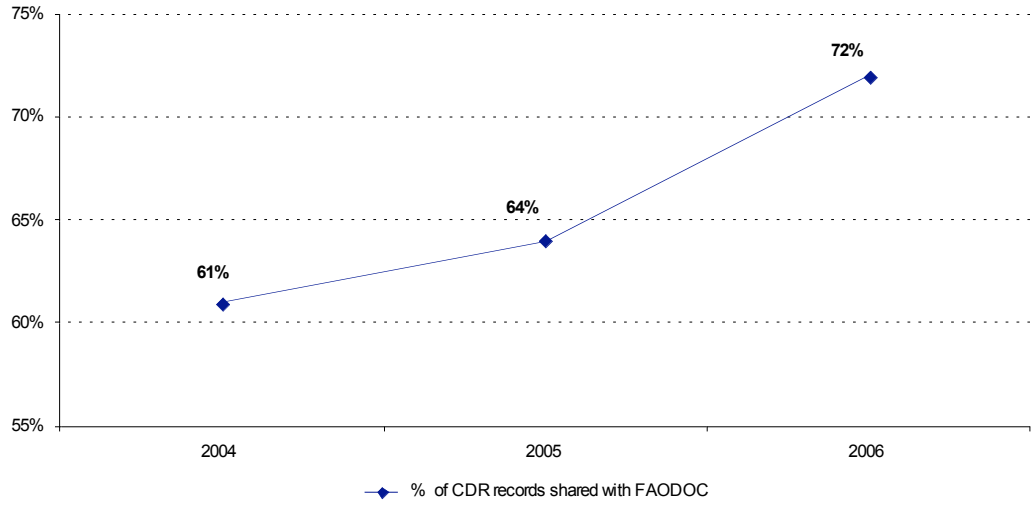

Figure 3: Percentage of the EIMS-CDR records catalogued in the FAODOC

In the EIMS-CDR, each record corresponds to one document (e.g., a book or a meeting report). The FAODOC catalogues documents and their analytics (e.g., a document is considered a book and the analytics are its chapters). Therefore, a book can have more than one record. The one-to-many relationship of records will be taken into consideration when merging data from the two databases.

The content of the two databases partially overlap, resulting in duplicate bibliographic records. The percentage of the EIMS-CDR full text documents linked from the FAODOC has increased over time (Figure 3): 72 percent of all records created in 2006 in the EIMS-CDR have been linked to from the FAODOC. This implies a duplication of effort (at metadata management level) and jeopardizes the dissemination and the maintenance of the FAO's institutional memory.

4 The Approach to Create the FAO Open Archive

The FAO Open Archive is based on the integration of the electronic publishing and the bibliographic cataloguing requirements. This merger requires the analysis of current workflows to detect similar procedures and reorganise them into a single coherent workflow. This process should focus on:

1. system architecture;
2. workflow procedure;
3. compliance with international data content standards; and
4. exposing metadata in a standardized way.

4.1 The New System Architecture

The architecture of the FAO Open Archive should integrate all features that are currently managed through the EIMS-CDR and the FAODOC. The FAODOC only manages the cataloguing process, but the FAO Open Archive must include the facility to deal with the reception of documents workflow, and improve the cataloguing module. The electronic publishing system is structured as a modular system where each module deals with a specific aspect of document publication. This approach will remain in the new architecture, integrated with new functionalities.

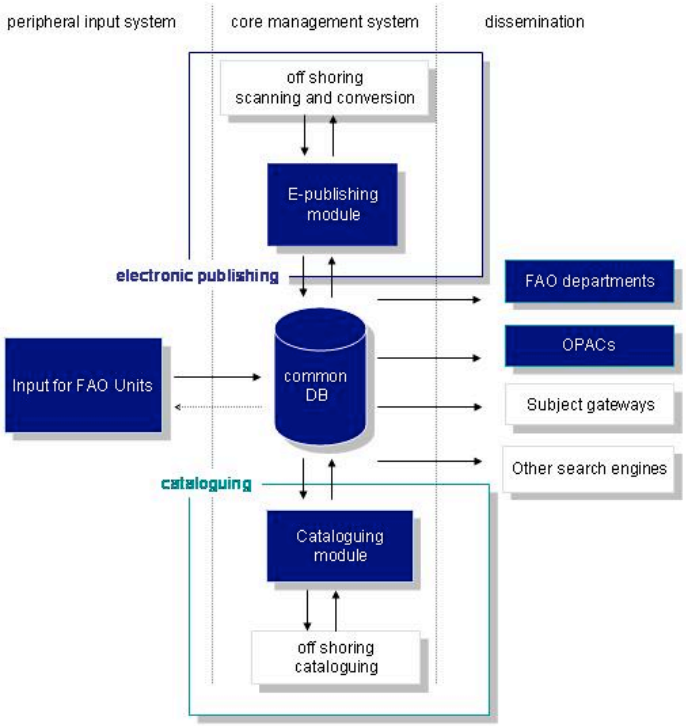

Figure 4: FAO Open Archive architecture

The FAO Open Archive architecture is detailed in Figure 4. The following elements define the architecture of the system:

1. integrated workflow; from left to right, the flow of information starts from the peripheral input system elements, passes through the core of the management system and to the dissemination interfaces;
2. common database; and
3. management of the two main functions of the FAO Open Archive; electronic publishing and cataloguing.

The objective of the system architecture is to manage all aspects of the electronic document life cycle. Electronic publishing and cataloguing will be managed through the same system and share the same database, e.g., from the document's creation, to its cataloguing, indexing and conversion to a suitable electronic format, to its dissemination on the Web.

Input for FAO units. This module will be used for data input and will be developed based on the current EIMS. FAO units now have individually customized EIMS interfaces. Each customization involves a basic internal workflow that can vary from one-step to multiple-step approval. FAO units are responsible for the introduction (and minimal description of documents) into the electronic publishing workflow. In the FAO Open Archive, FAO units will continue to provide data through a user-friendly system describing the document with a minimal set of metadata. With the FAO Open Archive, electronic publishing and cataloguing will share a common data entry point. The records that the FAO Open Archive will manage includes documents and multimedia files (photos, videos and audio) and non-FAO material (publications written in collaboration with FAO, yet FAO does not hold the copyright).

Electronic publishing. FAO will continue to publish documents online in electronic format. They will be managed through two modules:

- core module for electronic publishing – this module will be used to review the information from FAO units, based on EIMS, and to manage the conversion of full text documents into electronic formats (HTML, PDF, etc.); and
- scanning requests managing module – this module will be directly connected to the core module for electronic publishing and will be used to keep track of the work assigned to internal resources or of the work orders sent for scanning and/or conversion to external companies.

Cataloguing. FAO will offshore the cataloguing, using the minimal set of metadata and the full text provided by the FAO units. FAO cataloguers will check and validate the offshored records in order to guarantee the quality of the bibliographic description for the full text documents. Cataloguing will also be managed through two modules:

- core module for cataloguing – this module will be used to select records to be offshored for cataloguing and indexing and to check metadata quality. It will be used exclusively by cataloguers to manage the information to be released into the Open Archive; and
- cataloguing offshoring module – this module will be directly connected to the core module for cataloguing and will be used to manage the XML exports of data to be catalogued by external companies and to manage import and validation of offshored records.

4.2 Workflow Procedures

As well as the architecture, the workflow of the FAO Open Archive must integrate two main activities that so far have been conducted separately: electronic publishing and cataloguing. Figure 5 shows a top-down representation of the new workflow:

1. FAO units initiate a record by inserting a minimal set of metadata into the data input module. Only minimal information is requested to initiate a record: author, title, year and job number (a FAO unique identifier). The system verifies whether the job number exists in the database. A simple validation workflow within the peripheral input system will ensure that the records inserted are eligible for publication in the FAO Open Archive.

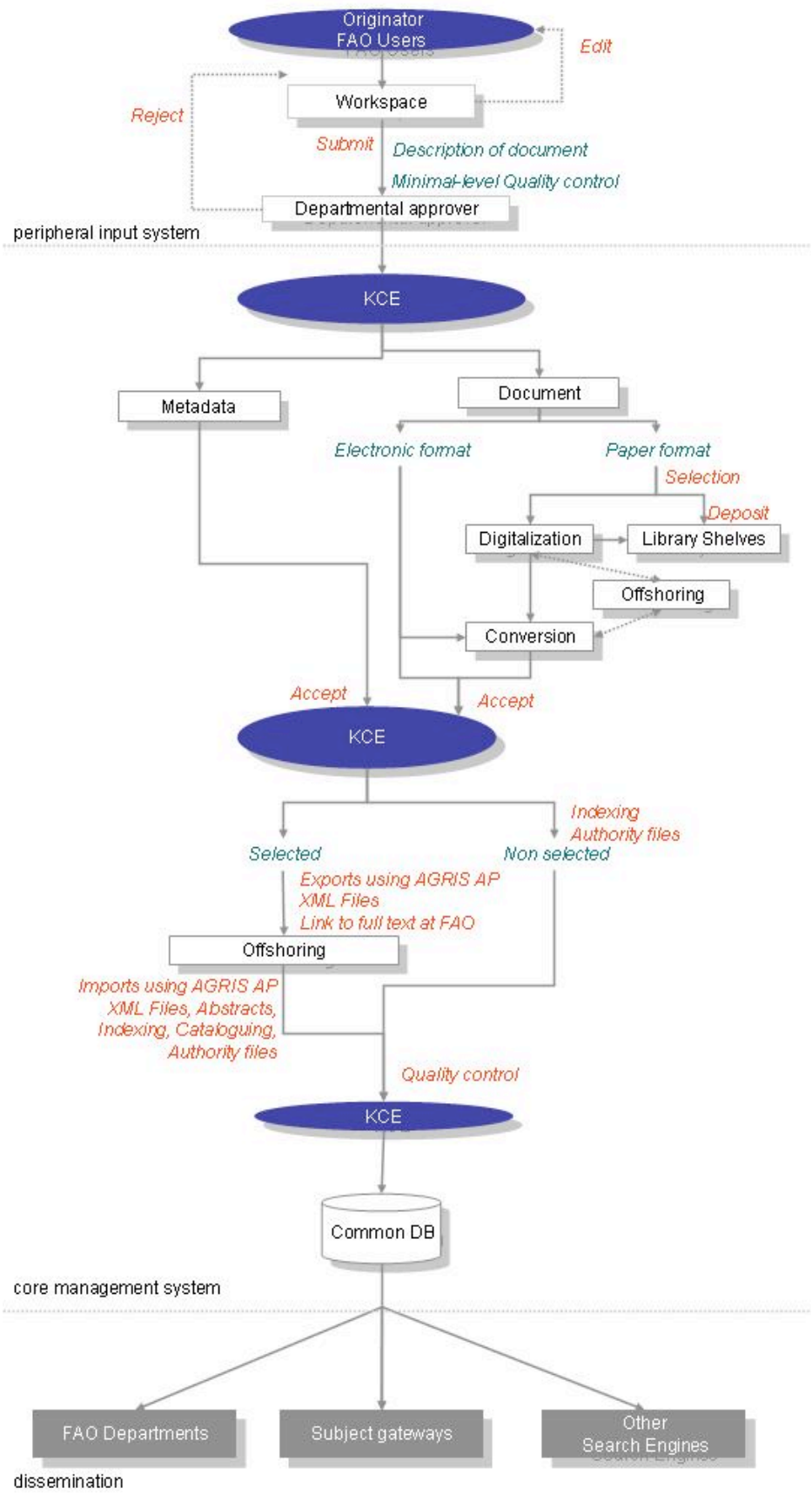

Figure 5: FAO Open Archive Workflow

2. The electronic publishing administration and the cataloguing administration are notified of the addition of a new record. They can take action simultaneously on the full text and the metadata of the records.

    2.1. If the document received is already in electronic format it requires validation and conversion to the most suitable format. This task can be carried out in-house or can be offshored. If the document needs digitalization then it is offshored for scanning.
    2.2. Using the minimal set of metadata in the system and the link to the full texts, the documents are catalogued and indexed by FAO and/or external cataloguers. The records that are selected for offshoring are exported using XML. When exported records are received from the external company they are imported into the system, checked and validated.
3. Validated records are disseminated through FAO Web sites. Moreover, search engines, services providers and digital libraries will harvest the records' metadata enhancing access to FAO documents.

4.3 Compliance with International Data Content Standards, ISBD

During the past few years, ISBD [11] has been identified as the standard most suitable for FAO. In April 2006, a study of the impact of changing FAO cataloguing rules recommended the adoption of ISBD rules:

> "*... recommend that FAO adopt the ISBD rules and build a system that will send and accept queries according to the OpenURL standard. In this way, FAO will build a system that will work with (interoperate with) other catalogues, while making FAO documents far more accessible to users. FAO, OCLC and other databases can create OpenURLs based on records that follow international guidelines and in this way, create an interoperable system* [12]".

ISBD rules are rigorous and exact. ISBD is based on the principles of adequate identification, searchability and consistency so that:

1. no two different documents can be confused with each other; and
2. the many details comprising a description, are presented in a uniform manner so that they can be interpreted without unnecessary ambiguity [13].

By applying the ISBD rules, FAO will not only enhance the international exchange of FAO records, but will also assist in the interpretation of records across languages, because ISBD records can be interpreted on a first level (identification of elements) by users of every language. This is because of the fixed order of ISBD records. Finally, ISBD is independent of any metadata format. In conclusion, ISBD rules are simple, exact, widely used and supported by the International Federation of Library Associations and Institutes (IFLA). ISBD will facilitate the interoperability with other institutions and/or services providers, as it is an international standard followed by many of the world's major libraries and bibliographic institutions.

One of the biggest challenges will be the handling of the legacy data; old records require re-cataloguing, e.g., titles need to be transcribed according to ISBD rules. A possible solution could be to import bibliographic records from databases that have already catalogued FAO documents, ignoring fields that are not relevant to FAO's needs and adding specific information already existing in FAO records, e.g., AGROVOC Thesaurus [14] descriptors. However, the legacy data can be updated, prioritizing those records which have the full text available and/or are accessed on a regular basis. The introduction of an additional code to distinguish old from new ISBD records is required.

The FAO units will introduce a minimal-level description based on ISBD and the offshored and FAO cataloguers could then bring the records to full ISBD level.

4.4 Exposing Metadata in a Standardised Way

This is a very important issue, and it has been addressed successfully by the Open Archives Initiative (OAI). Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) is a simple protocol that allows data providers to expose their metadata for harvesting to services providers. The FAO Open Archive will be OAI compliant, so the FAO metadata can be harvested by any services providers and/or digital libraries.

The concept of OAI-PMH can be applied to a wide range of digital materials, e.g. images, audio or videos. It is mandatory to expose metadata as Dublin Core. It is important to note that the protocol enables multiple metadata

formats. These alternative forms of metadata can be as rich as is necessary to describe content. During the last few years, FAO has made an intensive effort to promote the exchange of high-quality metadata within the AGRIS Network, an international initiative based on a collaborative network of institutions in agriculture and related subjects. The AGRIS AP is a metadata format that facilitates sharing of metadata across different information systems. It is a metadata schema which uses elements from metadata standards such as Dublin Core (DC), Australian Government Locator Service Metadata (AGLS) [15] and Agricultural Metadata Element Set (AgMES) [16] namespaces. The standard enhances the quality of the description of agricultural information resources, enabling greater processing possibilities by service providers. The AGRIS AP has proved to be a successful initiative, and as a result, the FAO Open Archive will be fully compliant with the AGRIS AP at export level.

In conclusion, exposing metadata will:

1. improve the retrieval of FAO documents from a large number of sources (e.g., portals, aggregators and services providers);
2. allow aggregators to detect FAO documents and thereby help to disseminate them; and
3. enhance the visibility and awareness of FAO's available resources.

5 Conclusions and Next Steps

This paper illustrates the first phase for the creation of the FAO Open Archive, focussing on finding a strategy to solve:

1. the duplication of efforts in creating and managing metadata; and
2. the lack of integration of electronic publishing and cataloguing.

The relevant findings from this first phase are:

- The FAODOC and the EIMS-CDR will use a common database and a workflow supported by a workflow management system. FAO will supply FAO bibliographic metadata together with the full text.
- The conversion of the FAODOC and the EIMS-CDR to the FAO Open Archive will facilitate the data input and maintenance of information. The FAO units will continue to be involved in the metadata creation process.
- The use of ISBD rules will simplify the creation of metadata. The legacy data will be updated to ISBD standards, prioritizing those records, which a) are accessed on a regular basis, and b) have the full text available to improve the effectiveness of the OpenURL protocol.
- The visibility and dissemination of FAO documents will be maximized by exposing content through OAI-PMH. The FAO Open Archive should have the ability to transfer and use information in a uniform and efficient manner across multiple organisations and information technology systems.

The creation of the FAO Open Archive will strengthen FAO's role as a knowledge dissemination organization. The following phase is related to the software implementation. The integration of open source software into FAO Open archive is still under evaluation.

Acknowledgements

We would like to thank Anne Aubert, Johannes Keizer, Giorgio Lanzarone, Romolo Tassone and Jim Weinheimer for their valuable contributions.

Notes and References

[1] FAO Constitution, Article I. http://www.fao.org/docrep/x1800e/x1800e01.htm#1 Last accessed in April 2007.

[2] Catalogue for Documents produced by FAO (FAODOC) http://www4.fao.org/faobib/index.html Last accessed in April 2007.

[3] Electronic Information Management Services (EIMS). http://www.fao.org/eims/ Last accessed in April 2007.

[4] Corporate Document Repository (CDR) http://www.fao.org/documents/ Last accessed in April 2007.

[5] The Knowledge Exchange & Capacity Building Division (KCE) of FAO is the responsible for all the above mentioned systems.

[6] AGRIS/CARIS Centre of Information Management for international agricultural research http://www.fao.org/Agris/ Last accessed in April 2007.

[7] International Standards for Bibliographic Description (ISBDs http://www.ifla.org/VI/3/nd1/isbdlist.htm Last accessed in April 2007.

[8] SALOKHE, G.; PASTORE, A.; RICHARDS, B.; WEATHERLEY, S.; AUBERT, A.; KEIZER, J.; NADEAU, A.; KATZ, S.; RUDGARD, S.; MANGSTL; ANTON. *FAO's role in Information Management and Dissemination – Challenges, Innovation, Success, Lessons Learned*. 2005. ftp://ftp.fao.org/docrep/fao/008/af238e/af238e00.pdf Last accessed in April 2007.

[9] This task involves the scanning and conversion of documents, corrections, modifications and the publication of HTML/PDF files.

[10] The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description http://www.fao.org/docrep/008/ae909e/ae909e00.htm Last accessed in April 2007.

[11] In 1969 the International Federation of Library Associations and Institutes (IFLA) created a general framework for the creation of standards to regularize the form and content of bibliographic descriptions (Byrum, J.D., "The Birth and Re-birth of the ISBDs: Process and Procedures for Creating and Revising the International Standard BibIiographic Descriptions". *IFLA journal*, Vol. 27, No. 1, 2001). The work resulted in the ISBD rules which specify the requirements for the description and identification of the most common types of resources that are likely to appear in library collections.

[12] WEINHEIMER, J. (2006). *Consequences of changing FAO cataloguing rules & format with ISBD/AACR2/MARC21: a report for the Food and Agriculture Organization of the United Nations*. Internal report.

[13] COETZEE, H. (2005). *Do we still need bibliographic standards in computer systems?* http://www.liasa.org.za/interest_groups/igbis/papers/IGBIS_WSJul04_Bib_Stds_Helena_Coetzee.doc Last accessed in April 2007.

[14] AGROVOC is a multilingual structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains. http://www.fao.org/aims/ag_intro.htm

[15] AGLS Metadata Standard http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html Last accessed in April 2007.

[16] Agricultural Metadata Element Set (AgMES) http://www.fao.org/aims/intro_meta.jsp Last accessed in April 2007.