Towards an architecture for open archive networks in Agricultural Sciences and Technology

Imma Subirats, Irene Onyancha, Gauri Salokhe, Johannes Keizer

Food and Agriculture Organization of the United Nations, Rome, Italy
{Imma.Subirats, Irene.Onyancha, Gauri.Salokhe, Johannes.Keizer}@fao.org

Abstract. The AGRIS Network is an international initiative based on a collaborative network of institutions, whose aim is to promote free access to information on science and technology in agriculture and related subjects. The paper illustrates how the Open Access (OA) and the Open Archive Initiative (OAI) models can be used within the AGRIS Network as a means of solving the problems of dissemination and exchange of agricultural research outputs. The lack of adequate information exchange possibilities between researchers in agricultural sciences and technology represents a significant weakness limiting their ability to properly address the issues of agricultural development. The OA model promotes the dissemination of research output at international, national and regional levels thus removing the restrictions placed by the traditional scientific publishing model. This paper presents the possibility to address the accessibility, availability and interoperability issues of exchanging agricultural research output.

Keywords: AGRIS, Agricultural Research and Technology, Open Access, Open Archive Initiative, Data Providers, Service Providers, AGRIS Application Profile, Knowledge Organization Systems, Interoperability

1 Introduction

AGRIS, operational since 1975, was a very early international initiative based on a collaborative network of institutions, whose aim was to build a common and freely accessible information system for science and technology in agriculture and related subjects. Until the late 1990s, AGRIS-related outputs mainly comprised a centralized bibliographical database and associated products. Since 2000, however, the AGRIS network has directed its efforts increasingly towards building up decentralized capacities in its participating resource centres, and empowering those centres to improve agricultural information management in their own institutions. One of the main undertakings of

AGRIS is to address issues related to accessing scientific and technological publications, and scholarly papers, but also and especially grey literature. Grey literature contributes significantly to agricultural research, especially in developing countries, but it is not yet widely accessible to the agricultural community.

Rapid technological changes have opened up new opportunities for sharing data, information and knowledge. Over the past few years, the communications world has witnessed several new developments, two of which are the Open Access (OA) and Open Archives Initiatives (OAI). These are complementary initiatives whose key objectives are to improve visibility and access to research publications, thus maximizing their impact. Open Access is described as "a free availability on the public Internet, permitting users to read, download, copy, distribute, print, search - or link to - the full texts of scholarly and scientific articles; crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without any financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" (BOAI, 2002). The Open Access model allows for the widest dissemination of research output, and maximum visibility, while, at the same time, removing the constraints common to traditional diffusion methods of scientific literature. It provides the means for researchers to avail themselves of full text content, using two paths: Open access publishing journals, which make articles openly accessible immediately upon publication, and Open Access archiving which allows authors "to deposit a digital document in a publicly accessible Web site, preferably an OAI-compliant open archive" (Self-Archiving FAQ, 2006). The Open Archives Initiative (OAI) develops and promotes interoperability solutions that aim to facilitate the efficient dissemination of content. The initiative is built on the principle that interoperability will federate the distributed open archives, thus encouraging the development of value-added services such as portals, subject gateways, and specialized search engines, with the overall benefit of increasing the visibility of the Open Archives. The ultimate objective is to overcome existing barriers to interoperability by using open archives for all digital materials (Van de Sompel & Lagoze, 2000).

The Open Access publishing model and the Open Archives initiative relate strongly to the AGRIS principles. By applying the former's concepts, the AGRIS community will not only build up the capacity of its resource centres to disseminate and share their research outputs, but will also bring accessibility and visibility to scholarly publications in agriculture. It goes without saying that the AGRIS network will also need to comply with the requirements and specific conditions outlined for the Open Archives.

This document defines and describes a high level architecture based on the Open Archives framework for the AGRIS community. The architecture comprises of three components: (i) creating content with agreed content descriptions standards, (ii) harvesting of this content using commons exchange standards, and (iii) providing value-added services to the users using the exchanged standard content. The paper is thus divided into three sections as follows: Section 3 discusses the overview of workflow architecture where the various components and the relationships between them are described. Section 4 presents the principal objective of this new architecture, which is based the Open Access model, and indicates the two actors namely data and service providers, their specific roles and their common formats, protocol of integration through metadata harvesting. Section 5 details the rational and use of exchange standards including metadata, Knowledge Organization Systems such as thesauri and common formats of integration which represent the building blocks of the semantic web. Finally, the conclusions in Section 6 provide the next steps in implementing the architecture.

2 Objectives

The AGRIS network will promote the Open Access publishing paradigm and apply OAI standards and methodologies since these latter significantly contribute to the realization of the long standing goal of AGRIS, namely to improve access to and exchange of information and knowledge in agricultural science and technology between developing countries. The specific objectives include increasing the visibility and accessibility of content including that which was not previously accessible and promoting common subject-oriented standards thus underpins the main principles of the AGRIS network. The

implementation of a new architecture using tools related to Open Access will greatly benefit the agricultural research community: both the visibility and readership of researchers will increase. Furthermore, the Open Archives offer easy access to documents for potential readers who would otherwise have limited access due to the cost of mainstream journals. The initiative enables linking of local and international research as well as providing a better picture of a country's agricultural research output. The new model provides an architecture in which communication between two tiers of partners can be defined: data providers, those who have metadata to expose, and service providers, who harvest the metadata. Communication between the two is enabled through an interoperability solution known as "metadata harvesting", which allows data providers to expose their metadata via an open interface to service providers, who use the metadata as a basis for value-added services. The quality of description of agricultural information will be based on the use of common subject- and network specific standards. Some of these standards include the AGRIS Application Profile (AGRIS AP) and Knowledge Organization Systems (KOS) such as the AGROVOC Thesaurus, elaborated further in Section 5. These will give the subject service providers in agricultural sciences better possibilities to build enhanced services from the metadata collected. Standards and tools will be promoted within the AGRIS network in order to guarantee the success of the new architecture.

3 Elements of the architecture: a new workflow

The proposed AGRIS network architecture, as further illustrated in Figure 1, provides the main elements and the foreseen dynamics of information flow in the network. Conceptually, the workflow has three principal activities: content management, exposing metadata and the provision of value-added services.

The content management activity describes the current process of capturing, describing and storing information using appropriate tools and methodologies. A list of appropriate standards, methods and software is available on the AGRIS and the AIMS[1] (Agricultural Information Management Standards) Web sites. A new element in the capturing process is that of self archiving of full text documents by

[1] AIMS Web site. http://www.fao.org/aims/. Last accessed in January 2007

researchers. The AGRIS centres will promote archiving of full text documents, either by the originator or a proxy within their institutes.

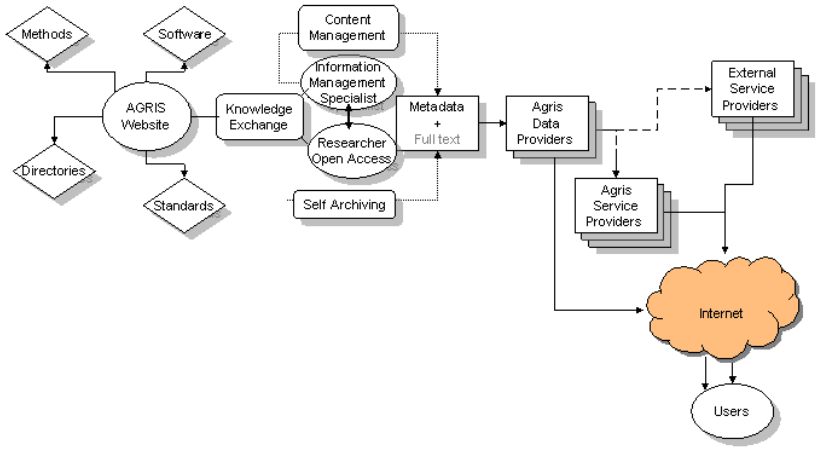

Figure 1. The new AGRIS Network Workflow

The second part of the workflow is to expose the metadata. This involves AGRIS partners becoming data providers and interoperating with service providers via the application of OAI standards and procedures. To guarantee visibility, the data providers will have to register and describe their collections in a registry available via the AGRIS web site. The AGRIS network intends to go beyond the exchange of simple Dublin Core metadata (a pre-requisite to participating in the OAI community), by recommending the use of the AGRIS AP - a richer exchange format for harvesting and exchanging agricultural content. The AGRIS AP and KOS are principal factors in the description of content. These allow the subject service providers to select and harvest quality data which ultimately improves the quality of their services. The last action of the workflow is that of subject services providers, who give added value to the metadata harvested, by offering platforms over which the researchers will be able to interact and share

information. The AGRIS network has established discipline-based communities which share common activities and services; these groups already form a basis on which the tasks and responsibilities of the service providers can be defined.

The service providers will also have the possibility to create specific subject gateways for different disciplines in the area of agricultural sciences. Furthermore, material held by the data providers will be also harvested by other service providers outside the AGRIS network. The overall picture of the new AGRIS network is that of a space where all the above described elements play an active role in the sharing of metadata and knowledge. The implementation of networks of open archives is the main element for the new architecture; it will allow the agricultural research community to both access and share its outputs.

4 The Data Providers: adopting the Open Access Model

Embracing the OA model, at the institutional level will mandate that the data providers to put mechanisms in place to support the implementation of digital repositories. These are open access archives that may include content such as already-published articles (post-prints), pre-published articles (pre-prints), theses, manuals, teaching materials or other documents that the authors or their institutions wish to make publicly available without financial or other access barriers. The technical infrastructure for setting up institutional repositories is readily available as seen from the variety of free open source software that adhere to the required standards and the limited costs associated to the set up. Set up of institutional ICT/M strategies and policies that include sound publishing and documentation workflow will also enhance the institutionalization of the Open Access model. Other issues to consider are the rights associated with author self-archiving of published articles. The ROMEO database (Rights Metadata for Open Archiving) available via the Securing a Hybrid Environment for Research Preservation and Access (SHERPA) Web site provides information on the publisher copyright policy and self-archiving. To facilitate this process, the benefits of OA should be clearly advocated and demonstrated amongst researchers as well as policy makers who can institute policies that favour and/or mandate OA. Libraries should champion this course within the institute by playing the important role of facilitating the adoption of OA through awareness campaigns to

educate its researchers and policy makers on the role and benefits of the model in promoting visibility and impact to the institution's research output.

4.1 Exposing metadata

Data providers in the new architecture are not comprised only of the existing AGRIS resource centres: any agricultural research and technology institution that wants to publish its output in the AGRIS network can register as a partner and exploit the common standards and technologies of the network. The data provider must be made aware that a repository or an open archive is not only a database with full text documents and their metadata. Figure 2 illustrates the various components of data provision: the content capturing process, where researchers can self-archive; cataloguing by information managers, and exposing metadata in the different formats. Three different types of data providing mechanisms are defined within the AGRIS network: Dynamic repository: a software layer which enables the repositories to be OAI-compliant and acts as an intermediary between open archives and the OAI-based service providers. The layer accepts OAI requests and generates dynamic responses; Static repository: used with small metadata collections. A static repository can be generated periodically by a script that extracts information from an existing database (Hochstenbach, Jerez & Van de Sompel, 2003) and Hosted repository: for repositories that cannot expose their own metadata on the Internet; other institutions can host and expose it for them. A data provider can expose different metadata formats, allowing them to be harvested by different service providers, thus giving them more outlet channels, and greater visibility. Unqualified Dublin Core is the basic metadata format for participation in the OAI community; however, it is not sufficient for the requirements of the agricultural scientific community. The AGRIS AP has been created specifically to improve the quality and interoperability of metadata in this domain. Data providers can offer both unqualified Dublin Core and the AGRIS AP in order to be visible to more search engines and subject service providers in the AGRIS network and beyond. The AGRIS Initiative will encourage the exchange of metadata using the AGRIS AP standard so as to guarantee the high quality description of agricultural information resources.

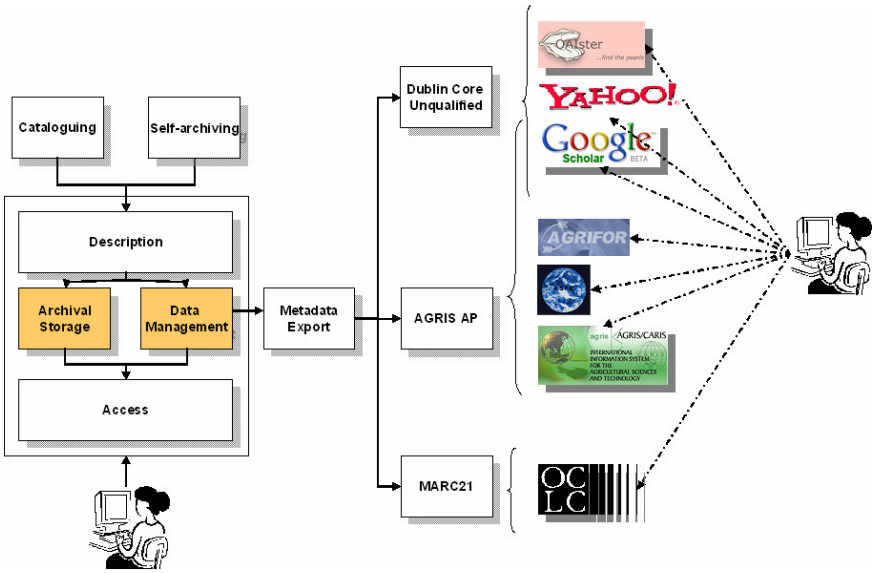

Figure 2. The Data Providers

5 Metadata Harvesting: creating the communication channel

Metadata harvesting is an interoperability solution to allow service providers to communicate with data providers. The interoperability solution in this case comprises uniform naming (standardized metadata), a common syntax, and a metadata harvesting protocol. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol developed with the objective of collecting the metadata of records found in open archives. OAI-PMH is based on a client-server architecture (service providers request metadata from data providers-open archives). Consequently, the service providers can build up their services using metadata from many different archives. During the past few years the OAI-PMH has become widely known as a way to establish a connection with distributed open archives. The main reason for its acceptance is the simplicity of its implementation. Also, several tools have been developed to allow for the creation, management, and interoperability of OAI compliance of the open archives. Uniform naming is achieved through the standardization of metadata semantics.. A common syntax, XML, is used for representing and transporting archive-specific metadata sets. Most current software is able to generate XML records.

6 The Service Providers: enhancing the access

The AGRIS network currently has several service providers, which include: the AGRIS database at FAO, the commercial AGRIS database at Ovid Technologies, the AGRI2000 service from Sistema de Informacion y Documentacion Agropecuaria de las Americas (SIDALC) , the Consultative Group on International Agricultural Research (CGIAR) virtual information centre, and many other smaller regional service providers. The new architecture delineates a much easier and more efficient environment to set up specific services. AGRIS Services Providers harvest, from more than one open archive depending on the scope and nature of their collections, records that contain only metadata, not full text documents. The functionality of a service provider to successfully harvest data is based on the provider's metadata schemas, metadata collection and open archive update frequency, the extent of the availability of the open archives, conditions of use of records, mention of source, and link from service provider to data provider. In 2003, Foulonneau and Dawson (Foulonneau & Dawson, 2003) outlined the key points to consider in the implementation of a service provider, most of which are already fulfilled in the AGRIS network as follows: Data collection & analysis: most of the AGRIS members are already collaborating in the AGRIS database; Common metadata schema: the AGRIS AP is the established metadata standard for exchange of the agricultural information; Granularity of descriptions: the AGRIS AP provides a qualified metadata set that enables richer description of agricultural documents. It has agricultural specific schemes; Metadata crosswalks: there are crosswalks from the AGRIS AP to other metadata schemas; Terminology issues: several KOS in agricultural science have been developed and used; Multilingual environment: one of the principal characteristics of the AGRIS network.

The harvested metadata can offer associated full texts, to which the service providers enable access. Different service providers are envisaged such as subject portals, digital libraries, gateways or web pages exposing only the metadata. Services that can be built using the harvested data include the following: data indexing; advanced searching and browsing; usage statistics; citation parsing; the possibility to make comments on documents; multi-language support; user authentication; aggregated news feeds; e-forums; automated email

notification alerts, i.e. Selective Dissemination of Information services (SDI); RSS news feeds on conferences, events and courses in different disciplines; tailored services (customization by the user).

7 Networking for Interoperability: The AGRIS Web site and AGRIS Open Archives Directory

The AGRIS Web site is a communication channel between the data and service providers and the specific user. It will be a gateway to various information resources including a repository for the outputs of the Task Forces, the AGRIS Open Archives directory, data management standards and tools, as well as other relevant information. The Web site plays an important role in the proposed AGRIS network architecture, and is consistent with the priority placed on advocacy. The Web site is a key tool in promoting the adoption of the broad vision of coherence in agricultural information in the context of the new AGRIS initiative.

The AGRIS Open Archives Directory is a platform for the registration of all the AGRIS open archives. It will be searchable on the AGRIS Web site, and used to identify all the data providers in the community. The directory will categorize and list the wide range of open archives using a set of metadata that will provide relevant information. It will be searchable by country, location, subject, metadata format, software, type, material included and other criteria.

8 Assuring quality in metadata creation

Content description through quality metadata creation and use of standard terminologies is the basis of efficient content management as well as the development of value added services. Metadata standards such as Dublin Core (DC) and the AGRIS Application Profile (AGRIS AP) provide mechanisms for sharing information in a standardized manner by recommending the use of common semantics and interoperable syntaxes. The OAI standards mandates the use of unqualified Dublin Core metadata schema for exposing metadata through the OAI-Protocol for Metadata Harvesting, however this is not sufficient for the requirements of the Agricultural Scientific Community. Use of unqualified Dublin Core inhibits the development of efficient value added services due to the poor quality and loss of data amongst other things. As mentioned before, the AGRIS AP, a richly defined standard, was created to improve the quality, and better

exchange of metadata within the agricultural science community. The qualification of the metadata element facilitates better aggregation and subsequent retrieval of content. Standardized metadata, in terms of syntax and semantics used, facilitates interoperability., however, poor quality metadata can mean that a resource is essentially invisible within a repository and remains unused (Barton, Currier & Hey, 2003). The more specialized subject service providers will use AGRIS AP as a basic metadata format for metadata harvesting.

To ensure a minimum level of quality, certain metadata fields and the use of relevant controlled vocabularies need to be agreed upon. The AGRIS AP achieves this by: mandating five required elements and promoting the use of agriculture-specific controlled vocabularies such as AGROVOC, CABI codes etc.

8.1 AGRIS Application Profile: fitting the agricultural requirements

Dublin Core is a metadata format whose structure is the result of an international consensus. The Dublin Core Metadata Element Set (DCMES) defines fifteen elements for simple resource description and discovery, all of which are recommended and none of which are mandatory. The DC has been extended with further optional elements, element qualifiers and vocabulary terms. Unqualified Dublin Core is the metadata format required for basic interoperability for OAI Protocol Metadata Harvesting. However, within subject areas and communities other metadata specifications may be required; for example, it may be necessary to describe resources with complex structures and in a specialized way, as in the case of the AGRIS network (Liang, Salokhe, Sini & Keizer, 2006). The AGRIS AP was designed for the agricultural domain with its specific characteristics and requirements. It is a standard created specifically to enhance the description, exchange and subsequent retrieval of agricultural Document-Like Information Objects (DLIOs). The AGRIPS AP is a format that allows sharing of information across dispersed bibliographic systems. It provides further qualifications to the simple Dublin Core element set; local extensions which were considered necessary for the comprehensive description of agricultural information resources. The standard, developed in light of the new AGRIS vision, enhances the quality of the description of

agricultural information resources, enabling greater processing possibilities by service providers.

The use of AGRIS AP by the data providers within the AGRIS Community offers certain advantages to the service providers: defines a richer interoperability layer to aid secondary and tertiary resource discovery through the use of a qualified element set; gives flexibility to provide more relevant query results through targeted searching; provides rich, qualified metadata; gives adequate information about the content of the resource through the use of agriculture specific qualifiers such as classifications schemes, thesauri, etc.; and provides the capability to browse resources by subject, by country, by year etc. Although to be "officially" OAI-compliant, data providers will have to be simple Dublin Core compliant, as far as the AGRIS network is concerned, where "subject services" are provided, it is recommended that data providers be also AGRIS AP-compliant.

8.2 Knowledge Organization Systems: aid to semantic navigation

Knowledge Organization Systems (KOSs) are organized knowledge structures, examples of which include: Authority files, Classification system/Schemes, Concept maps, Controlled lists, Dictionaries, Fact sheets, Glossaries, Ontologies, Subject headings, and Gazetteers. These resources fall along a continuum, according to the explicitness of their semantics and their amenability to machine interpretation. KOS are words or phrases taken from a standardized set that help to resolve two main issues: the problem of two or more words that can be used to mean the same concept, like Fishing vessels/Fishing boats or Health risks/Health hazards; and the problem of two or more words that have the same spelling but represent different concepts, e.g. vessel (blood)/ vessel (fishing) or Ling (a heath plant)/Ling (fish of the cod family). The use of KOSs allows for the creation of more specialized services such as: browsing by keywords, country of coverage, searching by type of Document (patent, books, etc.), and limiting search results to one or more specific years or language or semantic navigation within the result set based on keywords identified in controlled vocabularies. These services are enabled through the use of controlled vocabularies and their explicit mention in the data exchanged, a possibility provided by the AGRIS AP. The use of machine readable formats like XML and

standard vocabularies to exchange metadata are the building blocks for the future semantic web in the area of agriculture. They are the building blocks on which applications, that exploit the benefits of the semantic expressivity in bibliographic metadata, can be created to improve access to agricultural information.

9 Conclusions and next steps

This paper proposes the architecture for open archives within the AGRIS network. The proposed architecture combines the experience and history of the AGRIS network with the new scholarly Open Access publishing paradigm and the international Open Archive Initiative. Implementation of this architecture will lead to a very important space on the web into which thousands of data providers will feed their publications and on which many services will be based. This promises to be a major achievement, comparable only to other big projects like PubMed in the medical sector. Furthermore, much less investment will be necessary in the development of: common standards and tools, open source tools for data and metadata management for data providers, data schemas and vocabularies and sets of tools to ensure OAI compliance of existing systems. Much work has been done already in all these areas, and some of it can be leveraged for the AGRIS network, namely: the DC metadata schema and general OAI tools, the DC-based AGRIS Application profile, exchange vocabularies/knowledge organization systems like AGROVOC, the WebAGRIS suite of metadata management tools, the adaptation of DSpace to the specific needs of the AGRIS network and the development of an AGRIS open source (Lucene based) search engine for service providers.

Next steps include a consensus building for its acceptance by the major stakeholders of the AGRIS Network, followed by the planning of individual projects touching specific components of the architecture and, finally the roll out of individual projects in the different countries and regions of the network membership

References

[1] Second Consultation on Agricultural Information Management (2002) AGRIS - A strategy for an international network for information in agricultural sciences and technology within the WAICENT

Framework. http://www.fao.org/DOCREP/MEETING/005/AC502E.HTM Last accessed in January 2007.
[2] Budapest Open Access Initiative. http://www.soros.org/openaccess/read.shtml Last accessed in January 2007.
[3] Eprints: Supporting Open Access. Selfarchiving FAQ. http://www.eprints.org/openaccess/self-faq/#self-archiving Last accessed in January 2007.
[4] Van de Sompel, Herbert; Lagoze, Carl (2002) The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2). http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html Last accessed in January 2007.
[5] Agricultural Information Management (AIMS) Web site. http://www.fao.org/aims/ Last accessed in January 2007.
[6] Hochstenbach, Patrick; Jerez, Henry; Van de Sompel, Herbert (2003). The OAI-PMH Static Repository and Static Repository Gateway. JCDL 2003. http://public.lanl.gov/herbertv/papers/jcdl2003-submitted-draft.pdf Last accessed in January 2007.
[7] Foulonneau, Muriel; Dawson, David. (2003) Expert Report 3 – Open Archives Initiative – Protocol for Metadata Harvesting - Practices of cultural heritage actors. http://www.oaforum.org/otherfiles/oaf_d48_cser3_foullonneau.pdf Last accessed in January 2007.
[8] Barton, Jane; Currier, Sarah; Hey, Jessie M. N. (2003) Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications Washington (USA). http://eprints.rclis.org/archive/00001972/ Last accessed in January 2007.
[9] Berners-Lee, T. and Connolly, D. (1995). Hypertext Markup Language - 2.0 http://www.w3.org/MarkUp/html-spec/html-spec_5.html Last accessed in January 2007.
[10] Liang, Anita; Salokhe, Gauri; Sini, Margherita; Keizer, Johannes (2006) "Towards an infrastructure for semantic applications: Methodologies for semantic integration of heterogeneous resources" To be published in the Cataloging & Classification Quarterly.