A Dynamic Approach to make CDS/ISIS Databases Interoperable over Internet Using OAI Protocol

F. Jayakanth, K. Maly, M. Zubair, and L Aswath

Authors:

F. Jayakanth is a visiting Fulbright fellow at the Computer Science Department, Old Dominion University, Norfolk, Virginia, E-mail: franc@ncsi.iisc.ernet.in;
K. Maly is the Kaufman Professor and Chair at the Computer Science Department, Old Dominion University, Norfolk, Virginia, US. E-mail: maly@cs.odu.edu

M. Zubair, is a professor at the Computer Science Department, Old Dominion University, Norfolk, Virginia, US. E-mail: zubair@cs.odu.edu
Dr. L. Aswath is a professor at the Department of Library and Information Science Bangalore University, India. E-mail: lalitha_aswath@hotmail.com

Abstract
Research Paper

Purpose: A dynamic approach to make legacy databases like CDS/ISIS, interoperable with OAI-compliant digital libraries (DL).

Design/Methodology/Approach: There are many bibliographic databases that are being maintained using legacy database systems. CDS/ISIS is one such legacy database system. It was designed and developed specifically for handling textual data. Over the years, many databases have been developed using this package. There is, however, no mechanism supported by the package for seamless interoperability of databases.

The Open Archives Initiative (OAI) addresses the issue of interoperability by using a framework to facilitate the discovery of content stored in distributed archives or bibliographic databases through the use of the OAI Protocol for Metadata Harvesting. (OAI-PMH). The protocol is becoming a de facto standard for interoperability of DLs.

Many of the legacy database systems that are in use today, to the best of our knowledge, for various reasons, are not OAI-compliant. This makes it difficult for the legacy databases to share their metadata automatically.

There are two possible approaches to make legacy databases OAI-compliant – static and dynamic. In this paper, we discuss the dynamic approach to make CDS/ISIS databases OAI-compliant.

Findings: The dynamic approach is a simple way to make legacy databases OAI-compliant so that they become interoperable with other OAI-compliant DLs.

Research Limitations/Implication: Developing a general-purpose tool to make legacy databases OAI-compliant may be difficult as the driver software needed for database connectivity may not be easily available. Also, the structure of the databases are likely be

so varied that identifying the fields that make up the databases programmatically will be very challenging.

1. Introduction

CDS/ISIS was developed in the 1960s (at that time for mainframe computers) by the International Labour Organization in Switzerland. Since the 1970s, UNESCO has been maintaining and making further developments, including in 1985 a version to run on microcomputers (Micro CDS/ISIS). The Windows version was launched in 1995 (http://www.unesco.org/isis).

The advent of personal computers increased the popularity of this package. Number of institutions and organizations worldwide are using Micro CDS/ISIS for managing their bibliographic data. The popularity of the software package is evident from the fact that, upon request, UNESCO distributed more than 5,600 copies of Micro CDS/ISIS in 2004. There are thousands of databases, spread across the world, that are being maintained using this package. In the present day context of interoperability of DLs, it would be very useful to make CDS/ISIS databases interoperable with other OAI-compliant DLs. Also, there is a need to build a federation of domain or geographic specific databases, which we believe is necessary for faster and a better way of information dissemination. For example, a federation of all the OPACs (Online Public Access Catalogues) of a region or specific domains, which is accessible over the Internet, will enable users to go to that single site and search across all the catalogues that have been federated. In the past, one of the biggest obstacles for building a federated service has been a lack of standard for sharing metadata over Internet. Recent efforts, such as the Open Archive Initiative, specifically address this obstacle. To take advantage of OAI protocol, there is a need to make the CDS/ISIS databases OAI-compliant.

There are two possible approaches to make existing bibliographic systems OAI-compliant - static and dynamic. In the static approach, the existing records of a database are exported onto a file and the exported file is made OAI-compliant (Jayakanth et al., 2005). The major limitation of this approach is that the records will have to be exported whenever new records are added or any changes are made to the existing records in the database. The dynamic approach, however, will make the existing CDS/ISIS databases OAI-compliant in the real-time. In this approach, a software layer acts as an intermediary between CDS/ISIS databases and the OAI-based service providers. The software layer will accept OAI requests and interact with the databases to generate dynamic responses. This approach has the advantage that the response to the OAI protocol requests will always reflect the current state of database records.

2. Background

2. 1 CDS/ISIS

CDS/ISIS has been specifically designed to handle textual (bibliographic) data efficiently. It has a flat file structure representing data in non-tabular fashion. The data elements of a CDS/ISIS database can be repeatable and can have sub-fields as well. Every data element has a descriptive name and a unique tag number associated with it.

The formatting language of CDS/ISIS is quite powerful and is basically used for two purposes - defining templates for displaying/printing the data elements and for the generation of flexible indexes, which are based on inverted file structure. A good understanding of the formatting language is the key for its efficient usage.

The search language of CDS/ISIS is based on Boolean and proximity operators. It supports searching not only on the fields that are indexed but also on the fields that are not indexed.

To access a database system programmatically, either the Open DataBase Connectivity (ODBC) or the Java DataBase Connectivity (JDBC) drivers are needed. For CDS/ISIS databases, there are no standard ODBC or JDBC drivers available. This makes it difficult even to publish a CDS/ISIS database on the Web. However, a number of common gateway interface (CGI) based freeware packages are available to publish the CDS/ISIS databases on the Web, WWWISIS is one such tool being developed by BIREME, an organization based in Brazil (http://www.bireme.br/bvs/bireme/I/mision.htm).

2.2 WWWISIS – A server for CDS/ISIS databases in the Web environment

BIREME has placed WWWISIS up to version 3 in the public domain. It has developed a much more sophisticated package; known by the name wxis, from version 4 onwards. It is a gateway program specifically designed to work as a server for the CDS/ISIS databases in the Web environment (http://www.bireme.br/products/). To use wxis as the gateway for CDS/ISIS databases in the Web environment, a license has to be obtained from the developers, but it can be used on a stand-alone system without the license. It is driven by a XML-based scripting language called the IsisScript, and works both from the command prompt as well as a server in the Web environment.

BIREME has also developed a dynamic link library (DLL) called the ISISDLL for CDS/ISIS, which facilitates invoking CDS/ISIS-related functions from Windows-based programs. Since we wanted to develop a general purpose OAI interface, we decided against using ISISDLL library for database interaction. Instead, we have used wxis, which is available as freeware for both the Windows and the UNIX-based OS platforms.

2.3 The need for interoperability of CDS/ISIS databases

Interoperability is the ability of multiple systems, with different hardware and software platforms, data structures, and interfaces, to exchange data with minimal loss of content and functionality (http://www.niso.org/standards/resources/UnderstandingMetadata.pdf). Interoperability facilitates in building a federated system, which serves as a discovery tool for end-users.

There are many bibliographic systems that are being maintained using CDS/ISIS. To the best of our knowledge, there is no automatic means to share the metadata contained in the individual bibliographic systems. End-users will have to discover the existence of such systems and visit them individually.

Many of the DLs that are being built today are interoperable - thus facilitating sharing of metadata automatically to build federated systems. Many of the existing and emerging CDS/ISIS databases containing useful bibliographic information would like to expose their metadata and be part of such federated systems. Hence it would be very useful to make CDS/ISIS databases OAI-compliant.

2.4 Protocol for Harvesting Metadata – the OAI-PMH

The OAI-PMH is a low-barrier, non-intrusive protocol. It is an HTTP-based request-response transaction framework for communication between a harvester and an OAI-compliant data provider (Lynch, 2001). The data providers maintain OAI-compliant repositories, expose their metadata according to a metadata standard, and register their base URLs with OAI-based service providers. The service providers gather the metadata from the registered data providers and provide federated search service. The OAI-PMH supports multiple metadata standards, but the minimal requirement for the data providers is that they should expose their metadata at least in unqualified Dublin Core (DC) metadata standard (http://dublincore.org). The core document on the OAI-PMH providing technical details for implementers is available at the Open Archives Initiative Web site (http://www.openarchives.org/OAI/openarchivesprotocol.html).

2.5 Interaction between OAI data providers and a service provider

In a typical OAI-compliant systems (such as the software DSpace (http://www.dspace.org/) or EPrints.org (http://www.eprints.org/) for institutional repositories), the OAI interface of the system acts as an intermediary between a data provider and an OAI service provider. Data providers maintains OAI-complaint repositories and register their base URLs with OAI service providers such as at the ARC- Cross Archive Search Services at Old Dominion University (http://arc.cs.odu.edu) or OAIster at the University of Michigan (http://oaister.umdl.umich.edu/o/oaister/). OAIster harvests metadata once a week from the registered OAI data providers, builds a federated index, and provides search and browse functionalities on the harvested data. The interaction between OAI data providers and an OAI service provider is illustrated in Figure 1.

Take in Figure 1

Figure 1 – Interaction between OAI Data Providers and OAI Service Provider
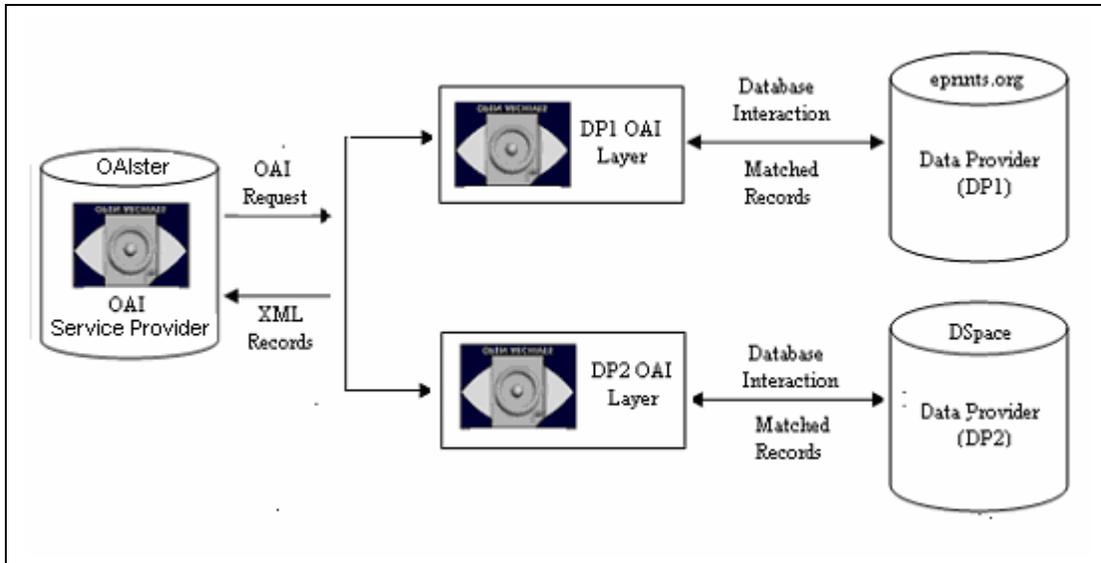
3. A dynamic approach to make CDS/ISIS OAI-compliant

3.1 Standardizing the CDS/ISIS databases

The fields that make up a CDS/ISIS database have descriptive names. Each field name is also associated with a unique numeric tag number in the range of 1-32,767. All the interactions with the databases are based on tag numbers and not on field names. Hence using tag numbers and sub-field delimiters that conforms to a standard like MARC (http://www.loc.gov/marc/umb/) would facilitate in identifying the different fields programmatically and it becomes feasible to convert the data elements to other metadata standards, if required.

Designing a general purpose OAI interface for CDS/ISIS databases that is able to interact with every available database is a rather difficult proposition as the tag numbers of the data elements may not be conforming to a standard. In such a scenario, it will not be possible to easily identify the fields programmatically. For example, consider the structure of a hypothetical database as shown in Table I.

Take in Table I

Table I – Structure of a hypothetical CDS/ISIS database

Tag	Name	Type	Repeatable	Sub-fields
10	Title	Alphanumeric	No	a
20	Auth	Alphanumeric	Yes	abcd
30	Abst	Alphanumeric	No	a
40	Publ	Alphanumeric	No	nlm

The tag numbers used in the hypothetical database, though valid, have been chosen quite arbitrarily. For the same set of data elements, it is quite valid to use a totally different set of tag numbers like 100, 110, 120, and field names like Ti, Au, Ab, Pu respectively. Identifying such variant tag numbers programmatically is very challenging. On the other hand, consider the structure of another hypothetical database as shown in the Table II.

Take in Table II

Table II – Structure of a hypothetical CDS/ISIS database conforming to MARC

Tag	Name	Type	Repeatable	Sub-fields
245	Title	Alphanumeric	No	a
100	Author	Alphanumeric	Yes	abcd
520	Abstract	Alphanumeric	No	a
260	Publisher	Alphanumeric	No	abc

The tag numbers and sub-field delimiters used in the above table conform to the MARC standard. Other databases that conform to MARC standard will use the same set of tag numbers and sub-field delimiters for their data elements representing title, author, abstract and publisher respectively. Hence the task of identifying the various fields programmatically becomes easier.

The tag numbers used in many of the existing CDS/ISIS databases are not likely to conform to any standard. Hence it becomes difficult to identify the fields programmatically. However, the reformatting FST feature of CDS/ISIS can be used to convert existing tag numbers and sub-field delimiters to conform to the MARC standard.

3.2 Reformatting FST feature of CDS/ISIS

This feature helps in converting the tag numbers and the sub-field delimiters of an existing database. For example, to change the tag numbers and the sub-field delimiters of a hypothetical database (as shown in table I), the existing records from the database are exported. A new database structure (as shown in table II) is defined. The tag numbers and the sub-field delimiters should conform to the MARC standard. The exported records can now be imported into the new database. As the tag numbers and the sub-field delimiters are different in the two databases, a reformatting FST file, as shown below, should be used to convert the tag numbers and sub-field delimiters of exported records to match the new database definition.

ID   IT   Data Extraction Format

245   0   mhl, (v10)
100   0   mhl, (v20)

```
520   0      mhl, (v30)
260   0      mhl, ("^a"v40^n, "^b"v40^l,"^c"v40^m)
```

The above reformatting FST file will translate the tag numbers 10,20,30, and 40 and sub-fields n, l, and m to tag numbers 245, 100, 520, and 260 and sub-fields a, b, and c respectively, during the import process. So all the existing CDS/ISIS databases, which are using non-standard tag numbers and sub-field delimiters can easily be converted to MARC standard and thus be able to expose their metadata for automatic extraction.

4. CDSOAI – The OAI interface for CDS/ISIS databases

The OAI-PMH protocol is becoming a de facto standard for interoperability of DLs. The data providers, who wish to take advantage such as a standard, will have to make their information/bibliographic systems OAI-compliant. This can be achieved by developing a software layer, which acts as an intermediary between the data providers and the OAI-based service providers. The CDSOAI is such a software layer, developed by us, which enables CDS/ISIS databases to be OAI-compliant.

CDSOAI (http://ncsi-net.ncsi.iisc.ernet.in/pmwiki/Main/Tools) is a java servlet-based OAI interface for CDS/ISIS databases. It acts as an intermediary between a data provider (CDS/ISIS databases) and a service provider (harvester). It relies on wxis, in the command mode, to interact with the CDS/ISIS databases. If the fields that make up a CDS/ISIS database conform to MARC standard, CDSOAI will serve as a general purpose tool for making CDS/ISIS databases OAI-complaint. A PHP-based tool was developed for the SciELO project (Marcondes, 2003) but, to the best of our knowledge, it is not a general purpose tool for making CDS/ISIS databases OAI-compliant.

When an OAI service provider sends an OAI protocol query to the servlet program (CDSOAI), the program parses the query and translates the query to a suitable search statement meaningful to wxis. It then builds a script file for wxis. This script file will have the name of the database, search statement corresponding to the service provider's request, and a template for displaying the data elements of the matched records. The control is then passed on to wxis, which will use the script file to interact with the database in the command prompt. The matched records are formatted as per the specifications in the template and are passed on to the servlet program. The servlet program then wraps the records in XML, conforming to the OAI:DC schema and sends them to the harvester. The harvester normalizes the harvested metadata and maintains a central index for federated searching.

The interaction between CDS/ISIS databases and a service provider through the intermediation of CDSOAI is illustrated in Figure 2.

Take in Figure 2

Figure 2 – Interaction between Arc Harvester and CDS/ISIS Databases through CDSOAI
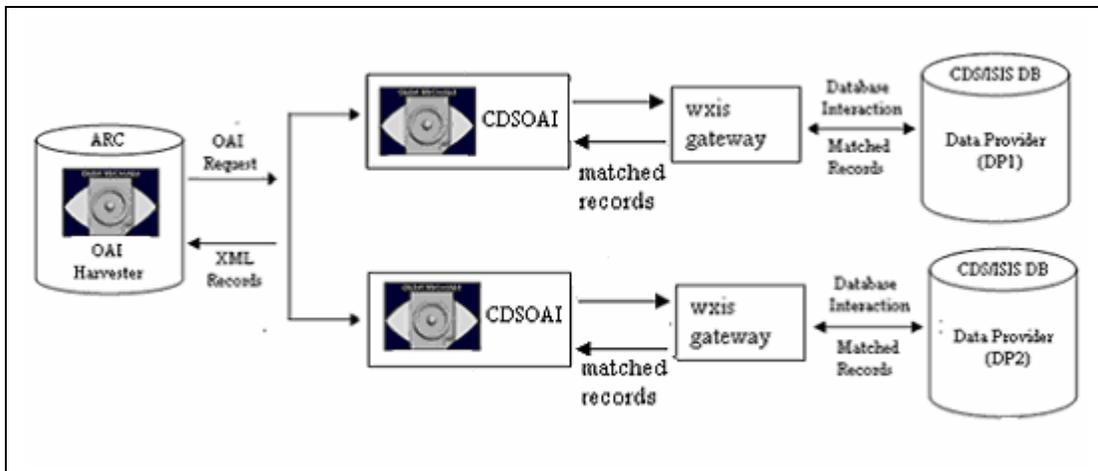
4.1 Federation of distributed CDS/ISIS databases

OAI-compliance basically facilitates in seamless harvesting of metadata from the data providers. CDSOAI transforms CDS/ISIS databases to be OAI-compliant thus facilitating harvesting of metadata by OAI service providers to build a federation. Such a federation serves as a discovery tool for the end-users.

Federation of OAI-compliant data providers can be easily implemented by using open source OAI service provider software like ARC (Liu, 2001), OAISTER (http://oaister.umdl.umich.edu/o/oaister/) or PKP harvester (http://pkp.sfu.ca/harvester/). OAI service providers harvest metadata from the registered OAI-compliant archives and build value added services on the harvested data.

To demonstrate harvesting of metadata from the CDS/ISIS databases using the CDSOAI software layer, ARC harvesting software was installed. Five CDS/ISIS databases were registered as data providers with the ARC. ARC successfully harvested metadata from the registered data providers and facilitates search and browse functionality on the harvested data. Figures 3 and 4 shows the screenshots of browse and search views of a prototype ARC harvester for CDS/ISIS databases
(http://etd.ncsi.iisc.ernet.in:8080/cds/index.html).

Take in Figure 3
        Figure 3 – Screenshot of Browse View of Federated CDS/ISIS Databases

        Figure 4 Screenshot of Search View of federated CDS/ISIS databases

        Take in Figure 3

Take in Figure 4

5. Conclusions

The CDSOAI can serve as a general purpose OAI interface for the CDS/ISIS databases provided the tag numbers and sub-fields of databases conform to MARC standard.

Databases not conforming to MARC standard can be easily converted using the reformatting FST feature of CDS/ISIS. It would be very handy, if a general purpose ODBC and JDBC drivers are developed for CDS/ISIS databases. This will make it convenient to publish CDS/ISIS databases on the Internet and also to develop tools like CDSOAI.

CDSOAI implements the OAI-PMH protocol fully thus facilitating automatic extraction of metadata from the CDS/ISIS databases. For the present, the interface supports mapping of extracted metadata elements to unqualified OAI:DC metadata schema. We intend to extend this to include support for other richer metadata schemata like qualified DC and MARC21. This will be useful for data providers who have much richer metadata set than simple DC.
.

Acknowledgements

References.

Jayakanth F., Maly K., Zubair M. and Aswath L. (2005), "Approaches to make
        CDS/ISIS databases interoperable with OAI-compliant digital libraries",
Program: electronic library and information systems, Vol. 39 No. 3, pp. 269-278

Liu X., Maly K. and Zubair M. (2001), "Arc - an OAI service provider for digital
        federation", *D-Lib Magazine*, Vol. 7 No. 4, Available at:
http://www.dlib.org/dlib/april01/liu/04liu.html

Lynch Clifford A. (2001), "Metadata harvesting and the Open Archives Initiative", *ARL
Bimonthly Report 217*. Available at: http://www.arl.org/newsltr/217/mhp.html

Marcondes, Carlos H. and Sayão, Luís F. (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:
http://www.dlib.org/dlib/march03/marcondes/03marcondes.html