Electronic Publishing '01 –
2001 in the Digital Publishing Odyssey
A. Hübler et al. (Eds.)
IOS Press, 2001

Converting heterogenous cultural catalogues and documents to XML - strategies and solutions of the COVAX project

Francisca Hernández(1), Peter Linde(2), Bob Mulrenin(3), Robin Yeates (4)

(1) Residencia de Estudiantes, Spain - fhc @interlink.es
(2)Blekinge Institute of Technology, Sweden – peter.linde@bth.se
(3) Salzburg Research Forschungsgesellschaft m.b.H., Austria – bob.mulrenin@ salzburgresearch.at
(4) London and South Eastern Library Region, United Kingdom - robin@viscount.org.uk

Abstract. COVAX (Contemporary Culture Virtual Archive in XML) [1] is an IST (Information Society Technology) funded project, launched as part of the IST first call, corresponding to key action 3 (Multimedia content and tools: cultural heritage and digital content) in the action line III.2.3 (Access to scientific and cultural heritage) under the European Community Fifth Framework for research and development. The main objective of COVAX is to test the use of XML to combine document descriptions and digitised surrogates of cultural documents to build a global system for search and retrieval, increasing accessibility via the Internet to electronic resources, regardless of their location. The project duration is 24 months. It started in January 2000 and the partners include content owners (memory institutions) and technological partners (developers: public RTD centres and private companies)[2]. COVAX's approach to achieving its objectives is based on the conversion of existing records to homogeneously-encoded document descriptions of bibliographic records, archive finding aids, museum records and catalogues, and electronic texts and on the application of XML (eXtensible Markup Language) and the various Document Type Definitions (DTDs) currently being used for library resource descriptions (MARC DTD), archives finding aids (EAD), museum materials (AMICO DTD) and electronic versions of cultural texts (TEIlite).

COVAX is designed to form a network of XML repositories as a distributed database. This will be accessed as a single database and will act as a meta-search engine, offering access to book references, finding aids, facsimile images, museum items, and other resources. COVAX is constructing a multilingual user interface to access such data. The project does not intend to create standards but to rely on the adoption of existing standards and concepts (XML, DTDs already in use, http...), using the Z39.50 protocol as a conceptual basis for communication between the multilingual user interface and the meta-search engine and Dublin Core Metadata Element Set elements as cross-domain access points. The conversion process has proved to be a crucial one in the COVAX-project and we therefore try to concentrate, in this article, on questions concerning our experiences of converting mainly library catalogues of different MARC or proprietary formats to XML.

1. COVAX OVERVIEW

The first stages of the COVAX (Contemporary Culture Virtual Archive in XML) project have been concerned with the definition of a COVAX architecture.

The purpose of COVAX is to analyse and draw up the technical solutions required to provide access through the Internet to homogeneously-encoded document descriptions of archive, library and museum collections based on the application of SGML/XML. The project will demonstrate its feasibility through a prototype containing a meaningful sample of all the different types of documents to build a global system for search and retrieval. It is based on the assumption that in libraries, archives and museums an enormous number of descriptions could be made available over the Internet by converting existing records or by creating new ones using specific SGML/XML DTD's.

1.1 User point of view

In the internet era, new challenges are constantly presenting themselves, and each new improvement in technique ends with a potential improvement in the functionality available to end-users. Sometimes these are founded on technical features such as speed, and sometimes on accessing new information resources. COVAX is expected to cover new challenges from this second perspective, transferring information resources to final users in a more adequate way, taking advantage of new XML technologies.The figure below gives an idea of what COVAX is expected to provide. End users, through the internet, will be able to access documentary information thanks to new XML paradigms.

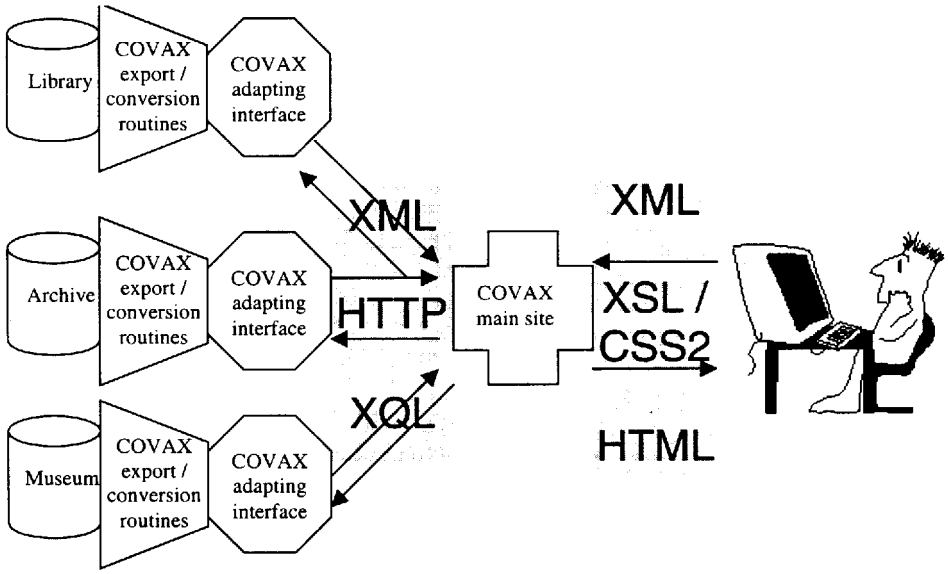

Fig. 1 - COVAX overview

The innovative aspect of COVAX is located in software layers that are invisible to the end user. The ability to obtain data from heterogeneous repositories, using emerging XML techniques, is where COVAX has to spend the maximum effort. Superficially, users will not notice any great difference from existing tools that access document databases.

1.2 Architecture design

The figure below shows, at the top level, the http client, which is the starting point of the COVAX system from the client side. At the next level, which might be called the servers level, at least two kinds of host will support COVAX: the COVAX web server, dealing with the main search and report capabilities; and the partner web server, specialising in the storage and indexing of information. Finally, at the bottom level, an off-line activity is shown, which is particularly important for the COVAX project: the conversion of current documentary data formatted for the XML databases.

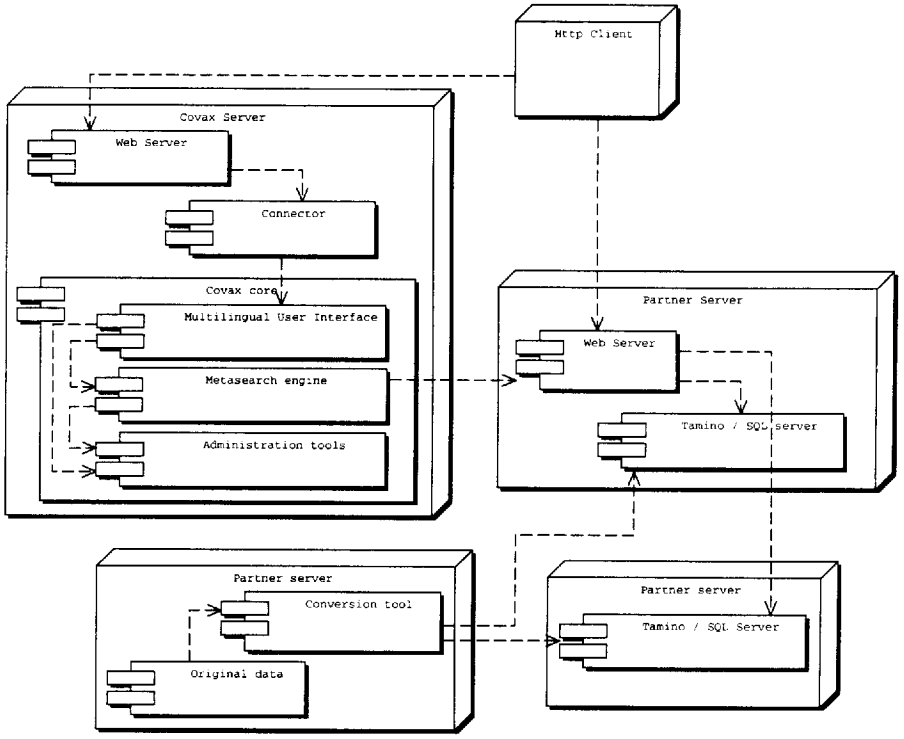

Fig. 2 - COVAX deployment diagram.

Physical architectural details as firewalls and external servers are not represented since the COVAX architecture is not dependent (it does not have to be dependent) on particular installations. It has to be adaptable for most typical environments, and this philosophy guides the design of the whole system. As currently defined, each partner system will be able to distribute the components of COVAX, without any requirement beyond accessibility to the internet.

1.3 Client

On the client side, there are two connections, one pointing to the main COVAX web server, and one pointing to the partner server. The user will access the main functionality of the system through the first connection. The second connection, directly to the partner server, is at present a possibility rather than a clearly defined function. We have envisaged the possibility of direct calls from downloaded HTML result pages to well-known sources. This connection is designed to allow this.

1.4 COVAX server

The COVAX server consists of three components: Multilingual User Interface, Metasearch Engine and Administration Tools. These provide the functionality expected from this system, comprising the core of the COVAX product:

Multilingual User Interface. Its role is "*developing the multilingual interfaces to access the main COVAX functionality*".

Metasearch Engine. It will "*develop the metasearch capabilities of COVAX, that will offer the following functionalities: in general, handle all the interaction with the COVAX individual library sites; and resolve COVAX-global queries*"

Administration Tools. Their role is "*development of the Administration Tools as required by the COVAX main site ("access server") being the middle-ware component as well as the user related and content provider related components*".

Control of the COVAX system resides in the Interface component. It is supposed in this approach, that each time the Interface is called, it will determine the functionality required (through service components such as the Metasearch Engine and Administrative Tools), and it will generate the corresponding required web page response. The Metasearch Engine will provide information needed by the Interface, and will obtain, from available resources, the information required by the user.

1.5 Partner server

On the partner server side, at least two valid configurations have to be taken into account, as shown in the figure. One or more nodes can be used to provide the partner functionality. Only one requirement has to be met: to provide an XML front end (capable of being queried in XPATH) accessible from the internet.
The conversion component is also shown, as an independent node. It makes available current information that is not directly accessible from an XML perspective. Records that have to be converted to XML format will be transformed in this component. It is represented as a different node for clarity, but no special architectural requirements regarding this have been detected.

2. CONVERSION SO FAR

The conversion process has proved crucial in the COVAX-project. We therefore think it valuable to share the problems we have encountered and the solutions we have implemented during the project so far - both general aspects and specific issues -for anyone wanting to convert documents and catalogues to XML for easier retrieval and interchange.

The first parts of the project were devoted to the design of the system, the conversion of existing records, and the software development for version 1 of the COVAX

prototype. The first design tasks were to define a coherent sample of records and documents from content owner partners; to analyse existing DTDs and select the appropriate ones; and to propose a common information structure.

A comprehensive dataset for the prototype was selected containing a wide variety of documents, descriptions, formats and databases: standard and non standard bibliographic records (including five different MARC formats), four different structures for archive and museum finding aids and information in six different languages (Catalan, Italian, English, German, Spanish, Swedish). The following excerpts from several content-owner partner returns gives an idea of the diversity of records and catalogues chosen for conversion:

Residencia de Estudiantes, Madrid - the union catalogue of bibliographic materials (monographs, serials and other) and archive documents held in 5 institutions (containing 75.000 records, mostly printed monographs, 1000 serials and 50 electronic publications, 100 videotapes and 100 sound recordings). Interchange formats: IBERMARC (USMARC – like). Archive finding aids: Over 36 personal and family archives from writers, scientists, painters, etc. related to Spanish contemporary culture. 3 archives from Spanish educational institutions. Some of the archives (10) are also described at catalogue level (20.000 records). In this case the format used is IBERMARC. During 2000 all the inventories will be converted to EAD and XML.

Angewandte InformationsTechnik Forschungsgesellschaft mbH, Graz - Swiss poster collection focussed on international contemporary posters from 1895 to date with about 200,000 records

Interchange format: USMARC/MARC21.

Universitat Oberta de Catalunya, Barcelona - Library catalogue, with a special emphasis on distance education, and virtual typology. Interchange format: CATMARC.

Blekinge Institute of Technology -The research document catalogue containing about 700 records, most of them with attached PDF-files for viewing the fulltext-document, administrated in a Lotus Notes Database. No standard Interchange format.

The library catalogue for the Naval Museum of Karlskrona containing about 2500 records. Interchange format: ISO 2709 Libris-Marc. Database: CDS-ISIS, version 1.0. Windows. Blekinge Institute of Technology is also providing about 30 full text XML-documents tagged according to the TEI-lite DTD. These documents are either museum-related or Swedish fiction.

LASER (London and South Eastern Library Region) - operates a union catalogue containing 4.27 million bibliographic record titles, in support of regional and inter-regional interlibrary loans and resource sharing. Interchange format: All records are held in UKMARC.

Salzburg Research – Museum object inventory and multimedia product archive. Both collections are based on a proprietary format. The museum inventory is from "Museum in der Fronfeste" leather collection stored in a Macintosh FileMaker database. FileMaker is often used by many small museums in the Salzburg area. The multimedia product archive is derived from the Europrix multimedia contest rights publication database. This database had already been converted to XML based upon a proprietary DTD and subsequently stored in an XML database server, Tamino from Software AG. Both collections include references to multimedia content, images and videos.

During the past 6 months all the content-owner partners have put a lot of energy into trying to convert these very different catalogue records and documents into XML format according to the specified DTDs.

3. DTDs selection

In the system definition, a crucial point was the selection of DTDs to be used in the conversion process. The decision was made on the basis of the State of the Art studies[3] developed in the early stages of the project. From the beginning, the possibility of creating specific DTDs for COVAX was rejected, and the DTDs created by institutions with

important standardisation capabilities were adopted (the above mentioned MARC DTD, EAD, AMICO DTD and TEIlite). The COVAX team has assumed that the influence, or even visibility of the project, was closely linked to the relevance of the standards adopted. The use of DTDs maintained by standardisation bodies also permits the use of a set of tools and procedures that will facilitate the adoption of XML by small or medium-sized Memory Institutions that do not have so many specialised staff or resources. All mentioned DTDs have been adapted to be converted from SGML to XML DTDs.

4. General Conversion issues

Once data structures to be used were defined, partners began conversion from original records, although there are differences between types of descriptions. Archival finding aids, museum descriptions and some of the electronic texts included in the prototype were created directly in EAD, AMICO or TEIlite. The main problem lies in the conversion of bibliographic records presented in up to five different MARC formats (IBERMARC, UNIMARC, UKMARC, CATMARC, LIBRIS-MARC). It has been necessary to plan different conversion processes: from original formats to MARC 21 (the basis of MARC DTD) using tools owned by the partners or by means of USEMARCON. In other cases, conversions were made directly, producing adequate code from original data structures using the MARC DTD. A free script provided by the Library of Congress, MARCCONV, was used to transform records to USMARC using the MARC DTD.

As explained in the previous section, unified DTDs have been selected to provide a common standard for transformation processes. These DTDs are the last structure that the records selected will show. Looking at the source data, one or two transformations are required.

We will use an example to explain the process of transformation from original data to the final XML documents. We consider the process for a non USMARC Bibliographic record. The correspondence with other possible formats is clear. The original non USMARC Bibliographic record has to undergo two transformations:

- To be semantically translated to USMARC.
- The semantically USMARC record, has to be converted to an XML structure.

These two steps can be performed in a single process, but the two 'logical' translations have to be carried out. Records in USMARC format will only require effort to translate them to the new XML format, without semantic considerations.

A small component diagram is shown below for a generic transformation process. The phase for adapting to USMARC format, is represented by one component representing the adaptation from any proprietary format to USMARC format. This proprietary format refers to any other non USMARC formats. The bottom component represents the structural transformation process in the case that the semantics are achieved. Three main formats have been considered as possible inputs for this component.

- ISO 2709, easily generated for most common bibliographic databases.
- Some installations can produce direct XML formats or even, raw data in XML format.
- Finally, proprietary formats that will require ad hoc developments.

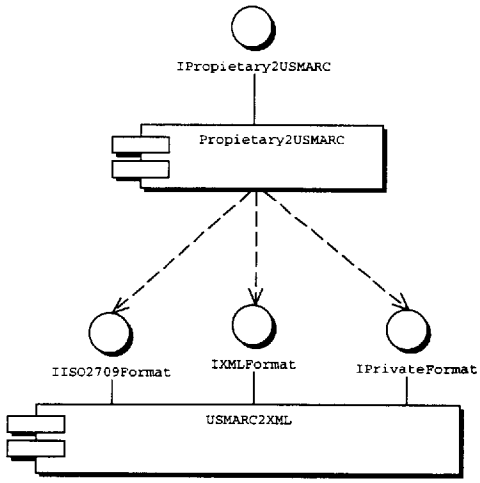

Fig. 3 - Migration component diagram

Interfaces represent the various detected input formats for this second step in the migration process. Once the record structure is achieved, it can be supported by records in ISO2709 format, XML format or any private proprietary format. For each input format, a different process has to be performed in order to obtain the common output: XML records with USMARC content.

4.1 Conversion types

As mentioned already, there are two ways to provide XML data. One is to have a front end translating on the fly some non XML data to an adequate XML format. The second is to store data in pure XML format, providing it in such an XML format when asked. We call them: adaptation and migration.

Adaptation.

Some partners in the COVAX consortium will investigate this approach. Taking advantage of records in relational databases, they will adapt the front end of the Data Base Management System to deliver results in adequate XML format.

Work on this kind of repository focuses on the appropriate configuration of front-end software, based on mappings between current data and the selected XML DTD.

Migration.

In this case, the data itself are stored in XML format. Work for this conversion involves the transformation of data from one structure to a new format. It requires a similar mapping to the previous method, but all transformations are made offline.

Data conversion will take place in several steps. As a result, converted data needs to be an XML representation of the underlying data. It might be adequate to transform the data into an internal representation (e.g.: XML with an proprietary DTD), which is then transformed using a target DTD. The conversion process can be detailed as shown in figure 4.

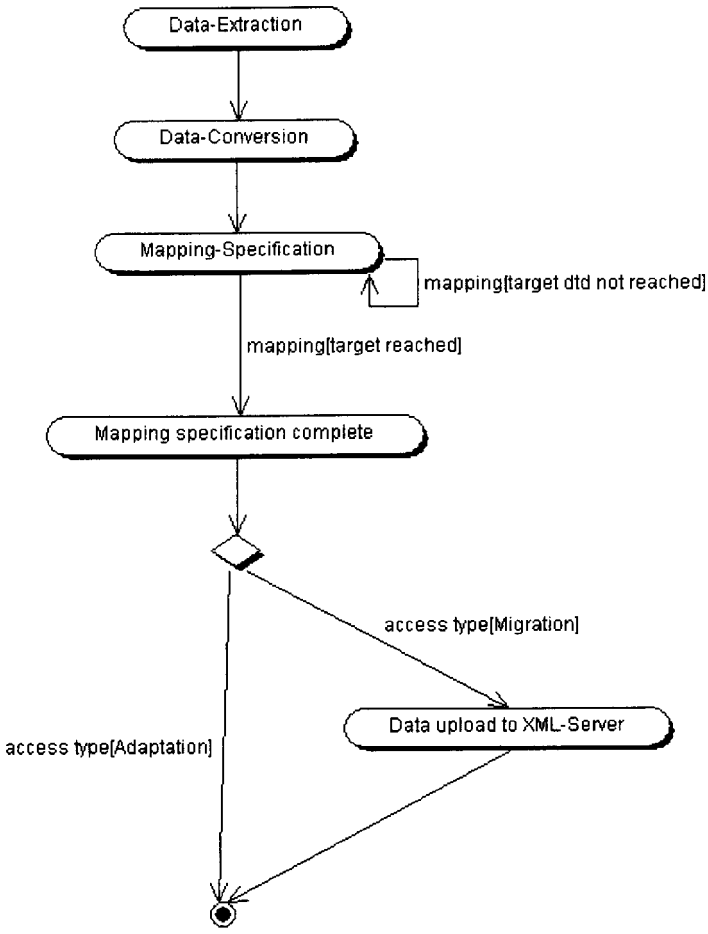

Fig. 4 – Data conversion process

4.2 Data-Extraction

The data needs to be available in an appropriate manner e.g. over a JAVA-Interface or in a text file. The data extraction activity therefore describes efforts to extract data from proprietary data stores such that other tools can use this data.

4.3 Data-Character-Conversion

Data-Character-Conversion takes place when the data does not support an appropriate character set. The mapping between different character sets is achieved in this step. Further discussion is needed to clarify if this step is part of the conversion tools or will be a service within the multilingual user interface.

4.4 Mapping-Specification

As the proprietary XML-representation may not be directly converted into the target format, multiple iterations of XML-mappings might be necessary.

4.5 Mapping-Specification Complete

When the target XML-representation is reached, it must be decided whether the conversion tools will implement an "adaptation" or "migration" strategy.

4.6 Software tested

Some efforts have been devoted to finding available tools on the web free of charge whenever possible to support migration or adaptation of records. Some success has been achieved. The most interesting tools tested are described.

Depending on the record format generated by the current information repository, different possibilities will be considered for the migration task. As previously defined, there are three different possibilities, according to the source format. Different tools can be considered for such formats.

4.7 Migration from ISO 2709

This is the most common format generated by document handling tools. Various tools on the web use this format for input:

Medlane Software. The Medlane project [MEDLANE] focuses on migrating ISO 2709 sources to XML records in Lane Medical Library. In addition to advancing XML as a representational format for document information, the Medlane work team has delivered a ISO 2709 to XML converter tool. There is a general tool to generate XML documents, from an ISO 2709 set of records, using any XML DTD as final structure. This implies that new DTDs can be used and applied during the transformation process. Another input represents the mapping between the original ISO 2709 contents and the final DTD structure. The tool is provided free of charge, and since August 2000, it can be used even in commercial systems.

Some experiments have been carried out using this tool, and minor successes have been obtained so far. We have completed some mapping between ISO 2709 input and the COVAX USMARC DTD, achieving partial success in the tests. We are in contact with a Medlane person responsible for help regarding this issue.

USEMARCON. User controlled Generic MARC converter [USEMARCON]. This is a tool converting ISO 2709 records between different formats:

UNIMARC to UKMARC (Bibliographic)
UKMARC to UNIMARC (Bibliographic)
UKMARC to UNIMARC (Authorities)
USMARC to UNIMARC (Bibliographic)
UNIMARC to USMARC (Bibliographic)

It has been tested with two sets of data: supplied test data and an ENEA sample of 100 records. The USEMARCON program installed and ran successfully with the supplied sample data. The ENEA sample was tested, but the default configuration is probably unsuitable for this data, as the program was unable to read past the first record. Further technical investigation is required to determine the exact cause of the problem.

BiblioML. BiblioML and AuthoritiesML [BiblioML] are XML-based formats for the interchange of UNIMARC bibliographic and authority records between applications. It is specific software for the translation of records from UNIMARC to some XML records following the structure described by two DTD's: biblioml_030.dtd and authoritiesml_030.dtd. We have not considered use of this tool due to the very specific nature of its approach.

MARCCONV. The Library of Congress provides a free tool [MARCCONV] that we have tested successfully. This tool translates MARC records from the ISO 2709 format to the published SGML DTD for bibliographic records. The pure XML requirements of the COVAX system have required some simple adaptations to the tested software, that have ended in the generation of XML documents (records) from several test records. More experiments are required with this tool before we can adopt it as a solution for translating bibliographic records in the context of COVAX.

The MARCCONV tool uses for reference a DTD (SGML) that is the SGML version of the XML DTD adopted by the COVAX consortium. This means that even if it is not possible (not easy at least) to adapt the software to new DTDs, the tool itself uses the same DTD structure as COVAX, or, rather, COVAX uses the same DTD as this tool.

5. Migration from XML

XSL (Extensible Stylesheet Language) is a standard (draft) [XSL] defining transformations of XML documents to other SGML formats. The typical application of such XSL definitions is the transformation of XML documents to HTML pages, in order to make the former more appropriate for user interfaces. The engine performing the transformation is called the XSLT (Extensible Stylesheet Language Transformations) processor.
The right combination of an XML source, for example UNIMARC XML records, and the appropriate XSL mapping the translation between UNIMARC XML to, for example, USMARC XML, would be considered another way of migrating records from one representation to another. The advantage of this method is the use of free and very common (each day more) tools. For instance software may include components like those proposed by [APACHE XML], such as Xalan (XSLT processor), Xerces (XML processor), and Cocoon (XML publishing and delivery components), and custom software, if desirable, to generate the XSL Stylesheet from templates and DTD mappings. The tools evaluated by COVAX partners are:

XML Translator Generator at IBM Alphaworks
This tool may be useful for converting XML documents based on one DTD to another DTD without using XSLT directly. This tool is a Java based XML processor (IBM XML Parser) that automates the translation of XML documents based on one DTD to another without the need for writing XSL scripts or program code. It uses examples of the same data in the two DTDs for generating a translator that can then be subsequently used for translation.
Tools to upload XML content to XML repositories (after conversion)
Following the conversion of data to XML, it will be necessary to bulk load the XML repository. Software AG [SAG] provides a new bulk loader, Java and Perl, with the Tamino 2.1 XML server
Adaptation from relational records
When the original data is located in relational databases, and no tools generating intermediate formats (ISO 2709, XML, ...) are available, we consider as a potential

solution direct reading of database records themselves. This can be done by ad hoc software or using existing tools. Tools considered are shown below:

XML Spy

[XML Spy] is a general tool useful for accessing and importing data from flat files or databases, including Microsoft Access. An XML document can be created when importing data and then transformed to the final XML document based on a COVAX DTD. Once an XML format is generated, if it does not conform to the COVAX DTD selected, it can be transformed using the Migration from XML tools.

5.1 XML migration XSLT combined with any data extraction tool

Data conversions from proprietary database formats or from non-standardized content are problematic. Mapping must be done to a DTD that may not be representative of all data in the source data. Furthermore, export tools and export formats may be completely non-standardized as well. What further complicates this situation is that the most current DTDs are not yet stable, especially for Museums. How does a content provider build a bridge to an XML repository?

5.2 Proprietary DTD and XML conversion document

One approach is to first build a proprietary DTD representing the database or a database view. The goal is to maintain an intermediate proprietary XML conversion document, although to create a new DTD is not a trivial task. We could, however, build our XML document without a DTD. The key is to export data with the database field names included within the source data. Using the database field names, we can build XML tags to encapsulate a particular field's data. One may visualise this process by using XML Spy to import a CSV (comma separated fields) file. Using this tool, we simple tell it to use the first row's labels as the column headers in the XML Spy table. Our XML tags are now defined. What we call column headings have become the XML elements of our intermediate XML document. XML Spy is one alternative for generating a proprietary XML document from a text file or database. Others exist, including specialized Perl scripts and Java programs.

Many database systems can generate CSV files with the first row containing the field names. FileMaker databases, found at many museums, can be exported as a *.mer file.

6. Why build an XML Conversion Document?

The conversion document represents the closest representation of the source data. From here, we have the ability to use XSLT stylesheets to transform our conversion document to an XML document based on a particular DTD. This DTD may be experimental or the most desirable at the moment. Many content providers will be participating in XML projects, each perhaps with standard or experimental DTDs. Our XML conversion document is our starting point.

6.1 Transformation of one XML document to another

We talked about an XML conversion document. Now we need an XSLT conversion stylesheet based on the target DTD (e.g. Amico-2in1, MARC DTD, etc) and our XML conversion document. For example, a simple sample XSLT stylesheet can be generated

from the target DTD by using TIBO's XML Authority. With this tool we create a sample XML document from the DTD and turn it into an XSL stylesheet that includes XSLT (transformation instructions). Our goal is to transform the proprietary XML conversion document into an XML document based on the standard DTD. In our case, the transformation from proprietary DTDs to the museum DTD, Amico-2in1, was used for three museums.

Lastly, our greatest difficulty in conversion is handling encoded data for which there is no standard, e.g. subject or classification coding scheme. During our conversion steps, should we retain the original non-standard coding scheme or attempt to map to another coding scheme? This is a major effort and requires a more sophisticated conversion tool with defined lookup tables that cross-map to a standard coding scheme.

What makes matters worse among heterogenous XML repositories is that most search engines must be aware of the various coding schemes found in the various DTDs. We can be sure that MARC coding schemes are known for each library, but encoding schemes for museum data vary widely. For now, we must wait for the data collection software used by museums and archives to achieve standardization in their metadata formats and encoding schemes.

7. SPECIFIC CONVERSION EXAMPLES

7.1 Converting IBERMARC bibliographic records

As stated earlier, one of the key points of the system definition has been the assumption of DTDs on which the corresponding databases are to be configured. During this definition process at Residencia de Estudiantes in Madrid, both existing initial database records and specific characteristics of the used formats and cataloguing policies have been carefully taken into account. In fact, bibliographic record conversion from IBERMARC to MARC DTD will need some pre-processing in order to make some field contents match MARC DTD definitions.

This is as a result of slight differences between IBERMARC and USMARC formats and of the evolution process of the MARC 21 format itself. This format is the result of two complementary processes: Integration format, which integrated into one single code the application variants of some fields in terms of the described material; and the unification process of formats in use by the major bibliographic agencies. Since 1999, MARC 21 has actually replaced USMARC, CANMARC (maintained by the National Library of Canada) and UKMARC (used for the British National Bibliography) formats and has thus become the most widespread format, used in far more records than any other MARC family format. At the same time there has been no updating or revision process of the IBERMARC bibliographic format (maintained by the National Library of Spain), resulting in the fact that these slight differences have been increased. In spite of this, these differences are not especially significant. They are stated here in order to show the profundity reached in the conversion process:

- Codes affecting positions 29, 30, 31 & 33 of control field 008 (fixed-length data elements) need to be modified.
- Contents of field 019 (Legal Deposit) are moved into field 017 (Copyright Information Number).

Repeatability of field 080 (Universal Decimal Classification) needs to be modified, because at MARC DTD, sub-field $a is repeatable, but the field is not. At IBERMARC it is just the reverse: the field is repeatable, but the sub-field is not.

- Contents of sub-field $j of field 650 (subject) are moved into sub-field $v.
- Contents of sub-field $v of field 852 (localization) are moved into $z (public note).

The figure below shows the process and an example of the conversion process results for a single record:

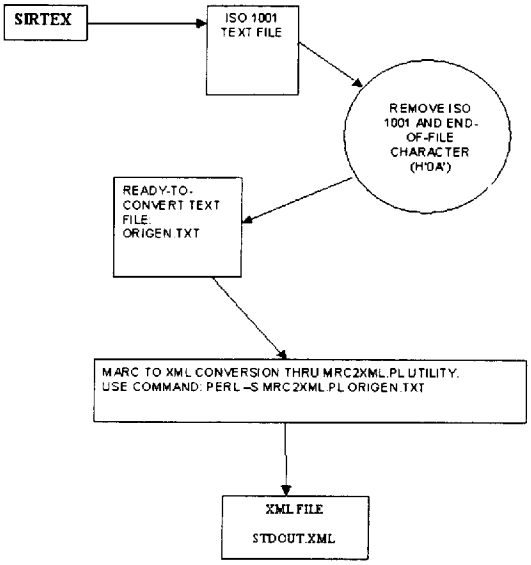

Fig. 5 XML Transformation process

Original Record:

```
00534nam  22001811  45
0010015000000050012000150080041000270350018000680400023000860800017001091
0000370012624500460016325000110020926000370022030000180025744000160027585
2006100291-AVC19990001281-19991122141-990701s1952     arg
spa d-  aMAXBI99000129-  aS(C)-FMAbspacCDM- 0a860-4"18/19"-1 aMenéndez
Pidal, Ramónd1869-1968-10aEstudios literarioscRamón Menéndez Pidal-  a7ª
ed.-0 aBuenos AiresbEspasa-Calpec1952-  a274 p.c18 cm- 0aAustralv28-40-
aS(C)-FMAbBMAjBMA/183zTiene algunas páginas subrayadas- 00607
```

Converted Record:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<mrcbfile>
<mrcb format-type="bd">
<mrcbldr-bd>
<mrcbldr-bd-05 value="n"/>
<mrcbldr-bd-06 value="a"/>
<mrcbldr-bd-07 value="m"/>
<mrcbldr-bd-08 value="blank"/>
<mrcbldr-bd-09 value="blank"/>
<mrcbldr-bd-17 value="1"/>
<mrcbldr-bd-18 value="blank"/>
<mrcbldr-bd-19 value="blank"/>
</mrcbldr-bd>
```

```
<mrcb-control-fields>
<mrcb001>AVC19990001281</mrcb001>
<mrcb005>19991122141</mrcb005>
<mrcb008-bk>
<mrcb008-bk-00-05 value="990701"/>
<mrcb008-bk-06 value="s"/>
<mrcb008-bk-07-10 value="1952"/>
<mrcb008-bk-11-14 value="blank"/>
<mrcb008-bk-15-17 value="arg"/>
<mrcb008-bk-18-21 value="blank"/>
<mrcb008-bk-22 value="blank"/>
<mrcb008-bk-23 value="blank"/>
<mrcb008-bk-24-27 value="blank"/>
<mrcb008-bk-28 value="blank"/>
<mrcb008-bk-29 value="blank"/>
<mrcb008-bk-30 value="blank"/>
<mrcb008-bk-31 value="blank"/>
<mrcb008-bk-32 value="blank"/>
<mrcb008-bk-33 value="blank"/>
<mrcb008-bk-34 value="blank"/>
<mrcb008-bk-35-37 value="spa"/>
<mrcb008-bk-38 value="blank"/>
<mrcb008-bk-39 value="d"/>
</mrcb008-bk>
</mrcb-control-fields>
<mrcb-numbers-and-codes>
<mrcb035 i1="i1-blank" i2="i2-blank">
<mrcb035-a>MAXBI99000129</mrcb035-a>
</mrcb035>
<mrcb040 i1="i1-blank" i2="i2-blank">
<mrcb040-a>S(C)-FMA</mrcb040-a>
<mrcb040-b>spa</mrcb040-b>
<mrcb040-c>CDM</mrcb040-c>
</mrcb040>
<mrcb080 i1="i1-blank" i2="i2-0">
<mrcb080-a>860-4&amp;quot;18/19&amp;quot;</mrcb080-a>
</mrcb080>
</mrcb-numbers-and-codes>
<mrcb-main-entry>
<mrcb100 i1="i1-1" i2="i2-blank">
<mrcb100-a>Menéndez Pidal, Ramón</mrcb100-a>
<mrcb100-d>1869-1968</mrcb100-d>
</mrcb100>
</mrcb-main-entry>
<mrcb-title-and-title-related>
<mrcb245 i1="i1-1" i2="i2-0">
<mrcb245-a>Estudios literarios</mrcb245-a>
<mrcb245-c>Ramón Menéndez Pidal</mrcb245-c>
</mrcb245>
</mrcb-title-and-title-related>
<mrcb-edition-imprint-etc>
```

```
<mrcb250 i1="i1-blank" i2="i2-blank">
<mrcb250-a>7ª ed.</mrcb250-a>
</mrcb250>
<mrcb260 i1="i1-0" i2="i2-blank">
<mrcb260-a>Buenos Aires</mrcb260-a>
<mrcb260-b>Espasa-Calpe</mrcb260-b>
<mrcb260-c>1952</mrcb260-c>
</mrcb260>
</mrcb-edition-imprint-etc>
<mrcb-physical-description>
<mrcb300 i1="i1-blank" i2="i2-blank">
<mrcb300-a>274 p.</mrcb300-a>
<mrcb300-c>18 cm</mrcb300-c>
</mrcb300>
</mrcb-physical-description>
<mrcb-series-statement>
<mrcb440 i1="i1-blank" i2="i2-0">
<mrcb440-a>Austral</mrcb440-a>
<mrcb440-v>28</mrcb440-v>
</mrcb440>
</mrcb-series-statement>
<mrcb-location>
<mrcb852 i1="i1-4" i2="i2-0">
<mrcb852-a>S(C)-FMA</mrcb852-a>
<mrcb852-b>BMA</mrcb852-b>
<mrcb852-j>BMA/183</mrcb852-j>
<mrcb852-z>Tiene algunas páginas subrayadas</mrcb852-z>
</mrcb852>
</mrcb-location>
</mrcb>
</mrcbfile>
```

7.2 Experiences converting small library and museum catalogues at BTH

As stated before, it is not unusual that small and medium-sized museums and archives keep their records in homemade MS Access-, FileMaker-, Lotus databases etc. Such was the case when the partners at Blekinge Institute of Technology during the conversion phase of the COVAX-project had to deal with several small catalogues in different formats for conversion to XML.

From our user network of museums, libraries and archives in the region we tried to pick a varied sample of catalogues and full-text documents. This included for example a municipal archival guide made in Access, a very simple database with only a few fields, that had to be converted into EAD-encoded XML records; the Research document database of Blekinge Institute of Technology created in Lotus Domino containing about 700 records in a no standard format; A museum object catalogue in FileMaker of about 20.000 records from Karlskrona Naval Museum; A sample of full-text documents from local archives, libraries and museums - exhibition catalogues, yearbooks, novels, maps, naval charts, to be converted from print to XML-markup according to the TEIlite DTD.

Unfortunately there are not very many relevant tools for converting material

like the above to XML. Most of our conversion effort has been a very time-consuming manual effort since we had very limited access to programmers.

Converting the Access archival guide, we simply used two methods: generating new records with proper EAD-tags out of data from database tables using Access' own report generator, and using the XML-spy-editor. XML-Spy has an "import database records" function that allows import of Access records and conversion of field-labels into the corresponding EAD-tag. This type of conversion was done as has been described above in the section named "Proprietary DTD and XML conversion document".

Choosing a DTD for the full text documents was never really any great dilemma since the collection that is being built at BTH and the samples needed for the COVAX project are contemporary cultural text-documents from museums, archives and libraries. The TEI (Text Encoding Initiative)is an international project to develop guidelines for the encoding of textual material in electronic form for research purposes. The TEI project have developed a set of very comprehensive DTDs that today are being used as more or less standards for marking up and interchanging cultural texts with SGML and XML. Very soon we found that one of these, the TEIlite DTD, suited our purpose of converting cultural heritage documents to XML, very well.

We used XML-Spy as the preferred editor when working with tagging the full-text according to the TEIlite DTD. We found it quite easy to work with XML-Spy since it has all the basic functions needed such as DTD support, the possibility to validate files and good import and export functions.

Tagging is very time consuming, but a straight forward thing as long as you are clear over what DTD, character set and elements to use in your documents.

Converting the Lotus Notes database, the main task was to map all the relevant fields to the MARC DTD. One problem here was of course the discrepancies in the semantics between the original research document catalogue and the MARC DTD. Another minor problem was the uneven quality of the catalogue record data.

When the mapping was done a Lotus script program was written which processed each record in the research document database and so generated a valid MARC XML-record for every record.

7.3 Converting Library Catalogs at LASER

LASER is involved in a number of projects with its members to digitize materials for the web, but its current database consists of bibliographic data. LASER operates a large union catalogue from which sample data needed to be extracted for COVAX. V3 Online (4) contains 4.27m bibliographic record titles, in support of regional and inter-regional interlibrary loans and resource sharing.

The selections for conversion were designed to ensure a number of identifiable 'collections' from a general purpose database as well as to reduce the number of records that need to be handled for the trial, so that COVAX content is reasonably balanced across sources.

LASER has experience of conversions from UKMARC to USMARC, and is aware of a number of difficulties with them. For example, we know we have a problem with converting UKMARC tag 248 (second level and subsequent level title) to USMARC, because it does not support this nesting of tags. We propose not to select records with 248 fields, or to delete that tag before output. In general a pragmatic approach has been taken during the project that represents the most replicable solution.

Other problems involve the selection of records which include a wide range of possible problem characters, especially special character such as the UK pound sign or rare characters that might cause loading and conversion problems later.

Since we planned to use standard existing conversion tools where possible, it was not deemed necessary to create full test records containing all possible characters and tags, and because of MARC flexibility, the possible range of complexity in records would be almost impossible to achieve by generating imaginary data anyway. Our emphasis was on producing realistic trial data that was shared via an intranet with partners from an early stage.

8. Data extraction and migration strategy

LASER has spent much effort outside the COVAX project to develop a transaction management system for ISO standard interloans based on its V3 database. In order to optimise development on the two separate systems, a migration strategy was preferred, so that work on COVAX did not interfere with work on the live systems. LASER was responsible for converting data from several MARC formats into USMARC, including UNIMARC, CATMARC and UKMARC. USEMARCON proved to require special configuration, and in general we preferred to adapt source systems to generate USMARC or UKMARC where possible. Samples were sent to LASER for validation and conversion to XML which was performed by using import procedures to convert records to Libpac internal format then export routines already in use to produce a range of USMARC products for publishers and libraries in Europe.

9. Data and character conversion

Having taken the decision to use USMARC and convert that to XML using Library of Congress MARC-XML Conversion Utilities (MARCCONV) with the mrcbfile.dtd (see http://lcweb.loc.gov/marc/marcsgml.html) we used XML Spy (downloadable free trial at http://www.xmlspy.com/) to validate the results and resolve problems. This is a complex and powerful editor that can be used to created DTDs and XML instances, but is also capable of validating XML against a declared DTD or XML schema. It is suitable for use by most library technical support staff. The main difficulty was deciding which character set encoding to use for XML output. UTF-8 was selected, although a number of problems remain with the handling of certain rare characters observed in data converted using MARCCONV, which we are currently working on at the time of writing.

10. Mapping specification

By choosing to use Libpac and Library of Congress tools, we avoided having to perform full mapping of UKMARC to MARC21 and from MARC21 to Dublin Core. Mapping is simply needed to ensure that all local tags required are included where necessary, such as our holdings information. We also needed to ensure that mandatory fields that will be required by COVAX system software are present. These included the MARC 040 tag (Cataloguing source), which for our purposes needs to contain a library code for each COVAX partner. We considered using Library of Congress codes for this (see the database and registration system at http://lcweb.loc.gov/marc/organizations/) but for administrative reasons chose to use simple codes during the initial pilot, in our case 'LASER'.

For some UOC and LASER tags, LASER bureau processing will strip tags from the source data during conversion to XML, by modifying the MARCCONV Perl script as required, or by pre-processing where this is more practical.

11. Validation

XML validation is performed using XML Spy, but even when data validates fully against the COVAX MARC DTD, we have found that there may be problems with characters appearing in the data. For example, the Tamino XML repository software in use at LASER for the project rejected UTF-8 records, and we then need to trace back to determine which software is validating wrongly. At the time of writing this is not clear.

12. Conclusion

It appears that conversion via migration of legacy MARC data to MARC21 and then XML can be done using available tools with minor modifications. Outstanding issues of character set encoding and holdings standards are likely to be resolved as we share experiences with partners within COVAX and subsequently beyond, and as the tools available are improved. We have already noticed that great strides forward have been taken by the Library of Congress since the project began, in such a way that European libraries can benefit. The recent announcement of the British Library's intention to change from UKMARC to use of MARC21 will further encourage adoption of MARC21 and the development of suitable conversion tools to handle legacy data such as ours. The next step is to mount live larger samples of our data (6,000 records) so that we can assess results display and bulk loading issues mid 2001.

13. References

[1]. *COVAX Web site* URL: http://www.covax.org/
[2]. The Project Co-ordinator is Residencia de Estudiantes (Spain) and the partners Angewandte InformationsTechnik Forschungsgesellschaft mbH. and Salzburg Research (Austria), Blekinge Tekniska Högskola (Blekinge Institute of Technology, Sweden), Software AG España, S.A., Universitat Oberta de Catalunya and Biblioteca de Menéndez Pelayo (Spain), LASER (London and South Eastern Library Region, UK), and ENEA (Italian National Agency for New Technology, Energy and the Environment, Italy)
[3]. State of the Art studies. *COVAX Web site* URL: http://www.covax.org/public_docum/p_documets.htm
[4]. V3 Online. URL: http://www.viscount.org.uk/products/v3online.html