CITIDEL Collection Building

Independent Study Report

Submitted by
Kunal Garach
kgarach@vt.edu

Under the guidance of
Dr. Edward A. Fox
fox@vt.edu

Department of Computer Science,
Virginia Polytechnic Institute and State University

Table of Contents

List of Figures

1. Introduction

The aim of this study is to facilitate the goals of the Computing and Information Technology Interactive Digital Educational Library (CITIDEL) [2] by increasing the number of collections available to it. This study will help in achieving this goal by focusing on the following collections:

1. Digital Bibliography and Library Project (DBLP)
2. ACM SIGCHI S.P.A.C.E. Student Exhibition
3. ACM SIGCHI CHI Tutorial Presentations
4. HCI Bibliography: Human-Computer Interaction Resources

1.1 Computing and Information Technology Interactive Digital Educational Library (CITIDEL)

On December 4, 2002 the United States National Science Foundation (NSF) officially launched the National Science Digital Library (NSDL) [3]. Composed of more than 100 interconnected projects, the NSDL is "a *center of innovation* in digital libraries as applied to education, and a *community center* for groups focused on digital-library-enabled science education." While initially launched by the NSF, the NSDL is clearly an international resource presenting materials and welcoming contributors and users from around the world.

NSDL is a portal to educational resources, realized through NSF support. It is an important source for science, technology, engineering and mathematics education, providing an abundance of materials for science education and learning.

Among the participating projects, a few are focused specifically on education in computing and information. In particular, the Computing and Information Technology Interactive Digital Educational Library (CITIDEL) seeks to connect educators in these areas with the resources and tools needed to guide the learning process for our students.

CITIDEL operates and maintains the "computing" portion of the digital library that includes information systems, computer science, information science, information technology, software engineering, computer engineering, and other computing-related fields. CITIDEL services the international computing community, and supports both English and Spanish language access to CITIDEL resources.

A consortium led by Hofstra University, The College of New Jersey, Pennsylvania State University, Villanova University, and Virginia Tech proposed to build CITIDEL as part of the Collections Track activities in the National STEM (Science, Technology, Engineering, and Mathematics) education Digital Library (NSDL). In particular, they establish, operate, and maintain a part of the NSDL that serves the computing education community in all its diversity and at all levels. This includes computer science, information systems, information science, software engineering, and computer engineering.

They help expand knowledge and skills regarding the creation and use of innovative online courseware for computing and information technology education. Even broader involvement results from submissions directed toward the paper, panel, tutorial and workshop portions of the program of conferences like FIE, ITiCSE, ISECON, OOPSLA, SIGCHI, SIGGraph, and SIGCSE.

Some of the features of CITIDEL are associated with the fact that it is a composite collection consisting of services layered over collections. It also encourages community participation. All the above features add up to “portal”.

Who can use CITIDEL?
CITIDEL can be used by anyone who teaches or is learning Computer Science or Information Technology. This includes but is not limited to professors and students at universities, K-12 teachers and students, workers in the IT industry, and people who train IT professionals

Resources in CITIDEL:
The sub-collections currently included in CITIDEL are:

?? CS Virtual History Museum
?? CSTC
?? NCSTRL
?? PlanetMath
?? ACM Digital Library

An important goal in building CITIDEL is to greatly increase the number of sub-collections, such as those mentioned above, included in it. This study focuses on helping CITIDEL to achieve that goal.

1.2 Background for this project

A number of important resources currently exist. They range from the extensive collections of computing literature in the digital libraries of ACM and IEEE-CS to lists of interesting web pages gathered and maintained by individuals. NSF and other funding organizations have supported creation of a wide array of resources, many of which would be of great value to teachers and learners if they were more widely known. Accordingly the proposal is to create, maintain, and operate CITIDEL as a portal (front-end) to all educational resources related to computing and information technology.

The Open Archives Initiative [4] Protocol for Metadata Harvesting provides an application-independent interoperability framework based on *metadata harvesting*. There are two classes of participants in the OAI-PMH framework:

?? **Data Providers** administer systems that support the OAI-PMH as a means of exposing metadata; and

?? **Service Providers** use metadata harvested via the OAI-PMH as a basis for building value-added services.

A repository is a network accessible server that can process the six OAI-PMH requests in the manner described in the standard document. A repository is managed by a data provider to expose metadata to harvesters.

Building upon work in the Open Archives Initiative, CITIDEL harvests metadata from all applicable repositories and provides integrated access and linking across all related collections.

The objective here is to make existing resources more easily accessible. CITIDEL is not a repository of resources; rather it is a site from which users can easily search and access all relevant resource repositories. These range in size and complexity over an enormous range. There is no point in trying to replicate what these repositories do. Instead, it focuses on making these existing repositories and lists more useful. Thus, considering the high-level architecture for the CITIDEL digital library, the emphasis is on the first two levels: User Portals and Digital Library Services.

CITIDEL does not try to replicate the above-mentioned lists, but aggregates their metadata through harvesting. An important aspect of the CITIDEL project is to determine which of these have ongoing support and are dependable sources of information for searchers and which have ceased to be reliable resources (and so should simply be integrated into CSTC or other collections).

Any study dealing with a heterogeneous set of collections is bound to face some difficulties due to the quality of the raw data available from these resources. Depending on the format (e.g., XML, HTML) and state (e.g., well defined, indistinct) of the source, it often is necessary to deal with the problems related to the quality of the raw data.

1.3 CITIDEL internal stream-of-records metadata format

The rationale for the CITIDEL internal stream-of-records metadata format [5] is to support services on top of data. With ‘one’ format for that data, the only need is to build ‘one’ set of services. The CITIDEL metadata format aims to provide this unified basis for CITIDEL's services. It is not for interchange at all, and is very tightly coupled to CITIDEL itself.

The CITIDEL internal stream-of-records metadata format can be described as a format that borrows from both the Dublin Core [6] and IMS metadata [7] formats. It also can be thought of as a superset of those portions of Dublin Core and IMS that are used for services.

Each record in the CITIDEL internal stream-of-records metadata format is made up of three optional elements: record, person, and links.

Record elements are further sub-divided into three elements: basic, LOM, and links. Elements under basic are concerned with general information (metadata) about the resource similar to the elements of Dublin Core. Elements under lom, which stands for Learning Object Model, are essentially similar to the IMS metadata’s LOM entries that use IEEE Learning Technologies Standards Committee (LTSC) Learning Object Model (LOM) [8] as its base.

Person entries are used to create person records that then can be used to link multiple works by the same author (system-wide) to a single person record. Link entries are used to generate system-wide cross-links between authors/creators and their works. An important feature of link elements is that they can appear inside as well as outside an actual record element.

A sample record in the CITIDEL internal stream-of-records metadata format is as shown below. In this case, it can be seen that the link element is defined inside the actual record. It very well could have been defined outside after the record ending tag.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE citidelstream SYSTEM "citidel_citidelstream_draft1.dtd">
<citidelstream>
<record>
      <basic>
            <identifier
            namespace="dblp">phdthesis.phd/Olken93</identifier>
            <title>Random Sampling from Databases</title>
            <date>1993</date>
            <author>Frank Olken</author>
            <institution>University of California at Berkeley
            </institution>
            <publisher>LBL Technical Report</publisher>
            <type>Text</type>
            <format>PHD Thesis</format>
            <source>DBLP</source>
```

```
    </basic>
    <lom>
        <resource_type>36</resource_type>   <!-- phd thesis -->
        <edlevel>4</edlevel>                <!-- graduate -->
        <edlevel>5</edlevel>                <!-- post-graduate -->
        <role>1</role>                      <!-- learner -->
        <role>5</role>                      <!-- teacher -->
        <source>8</source>                  <!-- DBLP -->
        <language>1</language>              <!-- english -->
        <offline>1</offline>                <!-- yes -->
    </lom>
    <links>
        <link type="is authored by">
        <target_title>Frank Olken</target_title>
        <annotation>Author of this item</annotation>
        </link>
        <link type="is published by">
        <target_title>LBL Technical Report Number: LBL-
        32883</target_title>
        <annotation>Publisher of this item</annotation>
        </link>
    </links>
</record>
</citidelstream>
```

1.4 Methodology

The collections that this study aims to deal with (see section 2) have their raw data in a variety of forms such as XML files, HTML web pages, and customized file formats. Each of these forms posed an interesting problem. It was necessary to study the raw data and determine the best way of approaching the problem. This was required because the metadata needed to be transformed such that CITIDEL could derive the maximum possible benefit from it.

The conversions as well as the method followed for each of the four collections have been discussed in detail in the form of case studies in the following sections. The discussion has been divided into sub-sections that follow the order:

?? Description of the resource: Provides a high level discussion of the resource itself and its raw data

?? Approach to conversion: Elaborates on the plan of action for that particular collection

?? Difficulties encountered in conversion: Discusses the various difficulties encountered while performing the conversion as sketched in the approach

?? Overcoming those difficulties: Discusses the means in which the difficulties were encountered and overcome

?? End result: Final result of the exercise

Each of these discussions is supplemented with a diagram showing how the transformations were carried out and, where possible, an example showing the raw as well as converted data.

2. Case Studies

2.1 Digital Bibliography and Library Project (DBLP)

Description of the resource

The DBLP server [9] provides bibliographic information on major computer science journals and proceedings. Initially the server was focused on Database Systems and Logic Programming; now it is being gradually expanded towards other fields of computer science. Thus DBLP is now read as Digital Bibliography and Library Project.

The server indexes more than 370,000 articles and contains several thousand links to home pages of computer scientists as of April 2003.

It is possible to search for authors and titles, and to browse the list of all conferences or the list of all journals indexed by DBLP. An alternative path is to go to the Subjects Page and to focus on areas of interest.

There are two levels of DBLP pages. The first involves **publication streams**. A publication stream enumerates the events of a conference series or the volumes of a journal. The pages contain links to the tables of contents (TOCs) of the proceedings or journal volumes. Bibliographic information on the proceedings, information on upcoming events, and pointers to web pages of the publishers are integrated into the DBLP publication stream pages.

Tables of Contents of proceedings or journals are the next level of the DBLP pages. The full list of authors, the title, and the page numbers are listed for each article. The name of each author is clickable: one may jump to a page that lists all registered publications of this person. If available, session titles and titles of special sections or issues are included into the TOC pages. The conference or journal name in the header line of the TOC pages is a back link to the publication stream level.

For a subset of publications DBLP provides additional information on **citation pages**. These are reachable by following the Electronic Edition links from the TOC pages. Electronic edition links with the remark "(link)" point to external citation pages provided by publishers like ACM or Springer. Other citation pages are part of DBLP and contain an abstract and information about an online or CD-ROM version of the article if available. For selected publications the list of references of the paper is reproduced on its citation page. References to publications known by DBLP are clickable.

DBLP is a bibliography server and not a document repository or delivery service. Many papers indexed by DBLP have been published only in printed journals or proceedings. One may find them at the local library or order them from the publisher or a commercial delivery service. They do not answer requests for papers. To access the electronic version of articles provided by publishers one may have to subscribe to their service. DBLP plays the role of an information broker that only refers you to a service provider.

Mirroring DBLP

The first task was to mirror the entire DBLP collection here at Virginia Tech. After successful mirroring, the metadata from the DBLP collection could be utilized and converted so as to be useful to CITIDEL.

The mirroring was successfully carried out by following the instructions listed on the DBLP website. The shell scripts provided on the website were customized for the local server here and the Mirror package [10] was used to perform the actual mirroring of the data files. The scripts used to accomplish the mirroring and subsequent periodic updates can be found online [1]. The mirror is presently hosted on one of the servers at Virginia Tech and can be located at http://crawler.cc.vt.edu/dblp/db

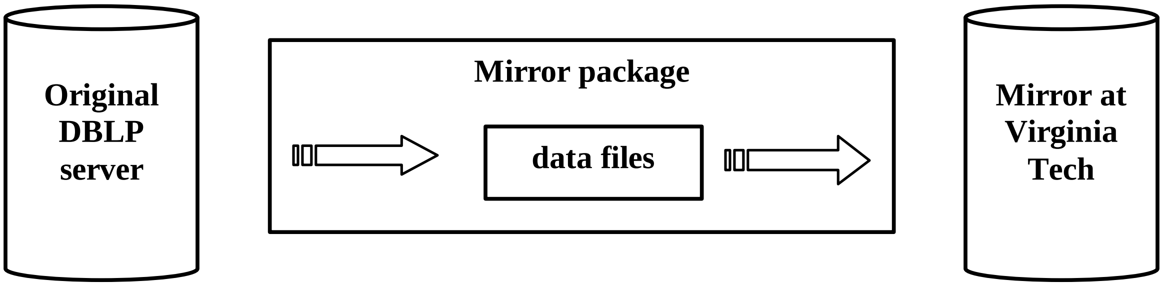

Figure 1: Mirroring process for DBLP

Following successful mirroring of the DBLP collection, the next step was utilizing its data.

DBLP data

For each paper contained in the DBLP collection a small file with the essential data is stored in a file system. The people in charge of DBLP reused the HTML parser and defined tags for the BibTeX record types and field names. Their bibliographic records look like

```
<article key="GottlobSR96">
      <author>Georg Gottlob</author>
      <author>Michael Schrefl</author>
      <author>Brigitte Röck</author>
      <title>Extending Object-Oriented Systems with Roles.</title>
      <pages>268-296</pages>
      <year>1996</year>
      <volume>14</volume>
      <journal>TOIS</journal>
      <number>3</number>
      <url>db/journals/tois/tois14.html#GottlobSR96</url>
</article>
```

Approach to conversion

All the metadata relating to the articles in the DBLP collection is available in a single file. This metadata is in the form of XML records that follow a customized DTD for the DBLP collection.

The approach followed was to convert all the metadata contained in the XML file from the DBLP XML record format to the CITIDEL internal stream-of-records metadata format. In this way, the metadata could be directly imported into CITIDEL.

Since the requirement was conversion of XML from one form to another, it was decided to perform this conversion using eXtensible Stylesheet Language Transformations (XSLT). The actual conversion was carried out using a publicly available Java conversion package, the ‘org.jdom’ package.

Difficulties encountered in conversion

The first and foremost difficulty involved with this conversion was the size of the original XML file. The file as of December 2002 measured ~128 MB in size. Attempting to perform any operation, such as reading records, required this huge file to be loaded, which could cause the entire server to crash.

The second difficulty encountered was to match the different tags in the DBLP XML format to those in the CITIDEL internal stream-of-records metadata format. The last challenge involved the actual conversion of the files, which included more than 320,000 records.

Overcoming these difficulties

The first step taken was to split the original XML file into multiple files so as to simplify any kind of operation on it. A program was written in Java that would accept the entire XML file and split it up into parts. The file was split such that each resulting file contained one DBLP XML record. The total number of split files numbered 320,009 giving an indication that there were 320,009 records in the DBLP collection.

Once the split files were obtained, an XSLT (eXtensible Stylesheet Language Transformation) was prepared to convert them into the CITIDEL internal stream-of-records metadata format. This involved studying the tags of the original XML and determining their equivalent in the CITIDEL internal stream-of-records metadata format.

This XSLT (eXtensible Stylesheet Language Transformation) was then applied to each of the source files to obtain the converted XML files. To do so, a Java program was written which utilized the publicly available org.jdom package. One of the applications of this package is to apply a given XSLT to an input XML file and generate a corresponding output XML file.

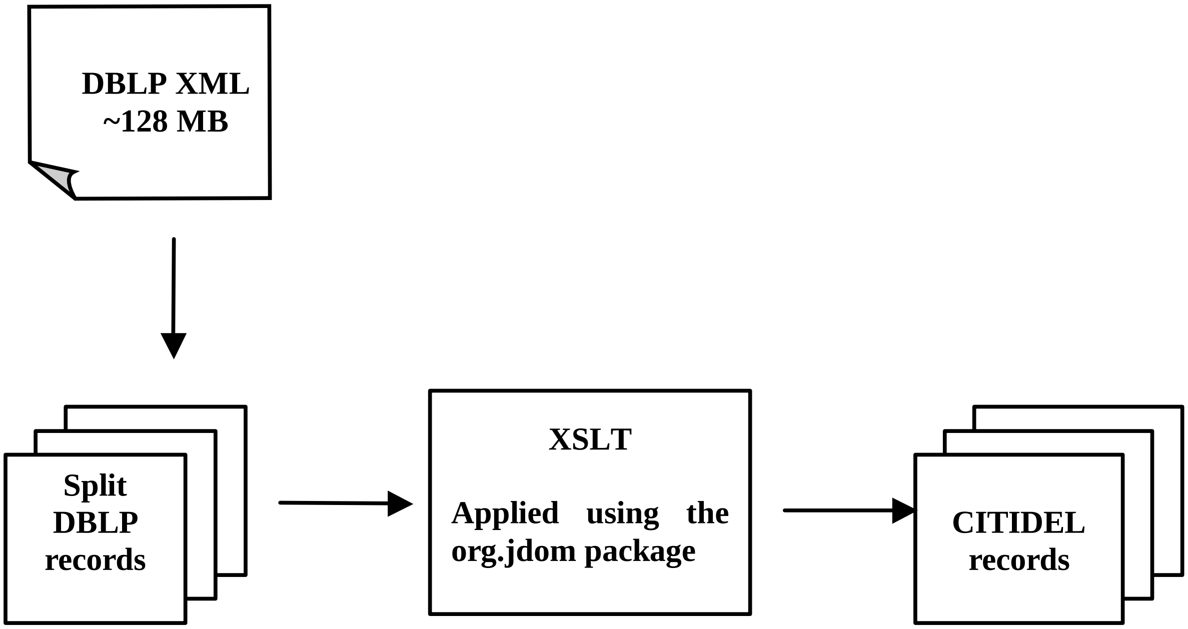

Figure 2: Conversion of DBLP data to CITIDEL format

org.jdom

JDOM is a very popular open source Java library, available from the JDOM website [11] that represents XML documents as a tree of Java objects denoting document elements. Although it duplicates much of the functionality provided by the W3C's DOM API, JDOM has a reputation of being both simpler and more efficient than DOM. The current release includes five packages: org.jdom, org.jdom.adapters, org.jdom.input, org.jdom.output, and org.jdom.transform. The org.jdom package contains the core classes that model an XML document. The transform package contains classes that allow a JDOM document tree to be used as either the input or output of a JAXP Transformer for applying XSL transformations. The principal differentiators between JDOM and DOM are JDOM's use of the Collections API and its use of concrete classes with a limited number of methods rather than interfaces with many methods.

End Result

?? The DBLP collection was successfully mirrored at Virginia Tech.
?? The metadata of the DBLP collection was successfully converted into the CITIDEL internal stream-of-records metadata format.

A sample record from the DBLP collection and the corresponding record in the CITIDEL internal stream-of-records metadata format is as shown below.

The original DBLP XML record is as shown below.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
      <article key="tr/gte/TR-0231-08-93-165">
            <ee>db/labs/gte/TR-0231-08-93-165.html</ee>
            <author>Frank Manola</author>
            <author>Sandra Heiler</author>
            <title>A 'RISC' Object Model for Object System
            Interoperation: Concepts and Applications.</title>
            <journal>GTE Laboratories Incorporated</journal>
            <volume>TR-0231-08-93-165</volume>
            <month>August</month>
            <year>1993</year>
            <url>db/labs/gte/index.html#TR-0231-08-93-165</url>
            <cdrom>GTE/MANO93c.pdf</cdrom>
      </article>
</dblp>
```

The same record after conversion to the CITIDEL internal stream-of-records metadata format is as follows.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE citidelstream SYSTEM "citidel_citidelstream_draft1.dtd">
<citidelstream>
<record>
      <basic>
            <identifier namespace="dblp">article.tr/gte/TR-0231-08-93-
            165</identifier>
            <title>A 'RISC' Object Model for Object System
            Interoperation: Concepts and Applications.</title>
            <date>August1993</date>
            <author>Frank Manola</author>
            <author>Sandra Heiler</author>
            <url>db/labs/gte/index.html#TR-0231-08-93-165</url>
            <url>EE: db/labs/gte/TR-0231-08-93-165.html</url>
            <type>Text</type>
            <format>Article</format>
            <source>DBLP</source>
      </basic>
      <lom>
            <resource_type>37</resource_type>   <!-- article -->
            <edlevel>4</edlevel>                <!-- graduate -->
            <edlevel>5</edlevel>                <!-- post-graduate -->
            <edlevel>6</edlevel>                <!-- professional -->
            <role>1</role>                      <!-- learner -->
            <role>5</role>                      <!-- researcher -->
            <source>8</source>                  <!-- DBLP -->
            <language>1</language>              <!-- english -->
            <offline>1</offline>                <!-- yes -->
      </lom>
      <links>
            <link type="is authored by">
            <target_title>Frank Manola</target_title>
            <annotation>Author of this item</annotation>
            </link>
            <link type="is authored by">
            <target_title>Sandra Heiler</target_title>
            <annotation>Author of this item</annotation>
            </link>
            <link type="is contained in">
            <target_title>GTE Laboratories Incorporated     Volume: TR-
            0231-08-93-165</target_title>
            <annotation>Journal this item is contained in</annotation>
            </link>
      </links>
</record>
</citidelstream>
```

2.2 ACM SIGGRAPH Student Poster and Animation Competition and Exhibition (S.P.A.C.E. Student Exhibition)

Description of the resource

ACM SIGGRAPH [12] is the Special Interest Group on Computer Graphics. ACM SIGGRAPH is extremely interested in supporting both Computer Graphics education and the use of Computer Graphics in education. The ACM SIGGRAPH Education Committee was established to accomplish this task.

The Education Committee has many different projects, involving volunteers from around the world in the areas of curriculum studies, resources for educators, and SIGGRAPH conference related activities. The ACM SIGGRAPH Education Committee also sponsors the computer graphics Student Poster Competition and Exhibition (S.P.A.C.E.). Selected projects are kept on exhibit at SIGGRAPH conferences. Selected posters from the S.P.A.C.E. entries may be included on the ACM SIGGRAPH Education Committee web site and CDs distributed at the conferences.

The interest in the S.P.A.C.E exhibit and competition has continued to grow from year to year, as has the quality of the entries. This juried competition provides an excellent opportunity for students working with computers to exhibit their creative work nationally and internationally. It is open to all students currently attending elementary or secondary schools, colleges, or universities.

The website for the S.P.A.C.E exhibit contains information concerning all student works submitted to the S.P.A.C.E competition from 1992 to 2002 [13]. Each of these years has its individual web page that contains information pertaining to that year’s competition. All the information on the web page needed to be converted into ‘collection-level’ and ‘item/record’ level metadata in the CITIDEL internal stream-of-records metadata format such that it can be directly utilized by CITIDEL.

Approach to conversion

The task in the case of the S.P.A.C.E. collection consisted of gathering information from web pages on the Internet and generating the metadata from this information. The first step involved analyzing these web pages to determine what metadata could be derived from them.

Since this was required to be an automated process, it was also essential to determine the common elements across pages/years and to form an outline for the data that should be extracted from the web pages and where this data would fit within the target format, the CITIDEL internal stream-of-records metadata format.

The information contained on the web page for each year of the competition had the following format: Jury and curator information was displayed first followed by information concerning the first, second, and third placed entries. This then was followed by a tabular description of all the other selected slides for that particular year.

Upon analysis of the web pages for the different years, it was determined that the best approach to this task was to generate a Perl script that would extract the information from the web page (essentially the metadata) and generate XML files which would contain this information in the CITIDEL internal stream-of-records metadata format. This Perl script would depend on the structure (HTML) layout of the page and require that this layout be consistent across the pages for different tutorials.

Difficulties encountered in conversion

The chief difficulty encountered was the layout of the metadata over the web page. Some of the data such as jury and curator information as well as honorable mentions (first, second, and third places) was in paragraph format while metadata regarding the rest of the slides was in tabular format.

Since the approach required parsing of the web page using a Perl script, it also required that the layout as described above be consistent across different pages/years of the competition. As it turned out, even though the tabular structure was more or less consistent across most of the years, there were a couple of instances where there was no structure at all.

Overcoming these difficulties

The final source code contained a Perl script that extracts the metadata by parsing the metadata contained on the target web page. This approach followed consisted of a combination of parsing based on the HTML structural layout of the web page and using the TableExtract module, obtained form the CPAN website [14], which helped in extracting information contained within the tabular structures on the web page.

The web page for each year was first parsed to generate collection level metadata for that particular year, including jury and curator information. This was accomplished by extraction of metadata based on the HTML structure of the web page. The next iteration was used to generate item level records for the honorable mentions of that particular year's competition. Again, extraction was based on the HTML structure of the web page. Finally, the last iteration involved using the HTML-TableExtract module to collect metadata at the item level about the other slides submitted to that year's collection.

As mentioned in the previous section, in some cases, where the structure did not follow the layout as required, some minimal manual intervention was performed which involved addition of tags to complete the source HTML and make it well formed.

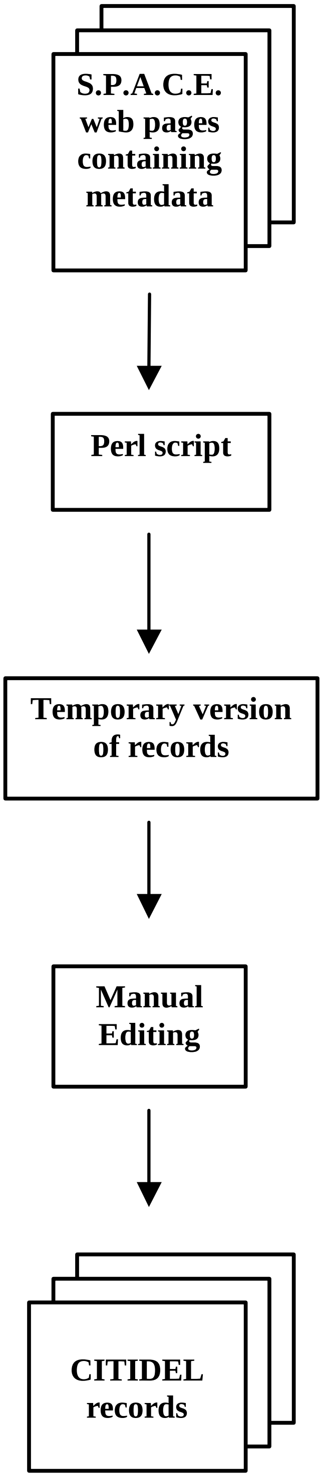
S.P.A.C.E. web pages containing metadata
Perl script
Temporary version of records
Manual Editing
CITIDEL records

Figure 3: Conversion of S.P.A.C.E. metadata to CITIDEL records

End Result

Parsing each of the web pages helped in collecting both collection level as well as item level metadata for each of the years of the S.P.A.C.E. competition. The corresponding XML files containing XML records in the CITIDEL internal stream-of-records metadata format were generated for each of the slides submitted to the competition.

Cross-links, which form an important component of the CITIDEL internal stream-of-records metadata format, between collection level records and the item level records, for each year, were generated. Person records for each of the creators/authors of the slides also were generated.

The only issue not resolved was the missing slide sets for the years 1998 and 1999 since information about these years was missing from the web site itself.

2.3 ACM SIGCHI CHI Tutorial Presentations

Description of the resource

ACM, the Association for Computing Machinery [15], is a major force in advancing the skills and knowledge of Information Technology (IT) professionals and students throughout the world. ACM serves as an umbrella organization offering its 78,000 members a variety of forums in order to fulfill its members' needs such as the delivery of cutting-edge technical information, the transfer of ideas from theory to practice, and opportunities for information exchange. Their services include providing high quality products, world-class journals, and magazines; dynamic special interest groups; numerous "main event" conferences; tutorials; workshops; local special interest groups and chapters; and electronic forums. ACM is the resource for lifelong learning in the rapidly changing IT field.

The scope of SIGCHI [16] consists of the study of the human-computer interaction process and includes research, design, development, and evaluation efforts for interactive computer systems. The focus of SIGCHI is on how people communicate and interact with a broadly defined range of computer systems. SIGCHI serves as a forum for the exchange of ideas among computer scientists, human factors scientists, psychologists, social scientists, system designers, and end users. Over 4,500 professionals work together toward common goals and objectives.

SIGCHI sponsors or co-sponsors many conferences related to human-computer interaction.

ACM SIGCHI brings together people working on the design, evaluation, implementation, and study of interactive computing systems for human use. ACM SIGCHI provides an international, interdisciplinary forum for the exchange of ideas about the field of human-computer interaction (HCI).

The annual CHI conference is the leading international forum for the exchange of ideas and information about human-computer interaction. Diverse members of the global HCI community meet at the CHI conference to share the excitement of discovery and invention, to make and strengthen professional relationships and friendships, and to tackle real world problems.

Tutorials are courses that offer extended interactions with expert instructors. The courses available at CHI 2000, CHI 2001 and CHI 2002 represent the leading edge of current practice and research in human-computer interaction.

Tutorials cover emerging technologies and markets, along with usability methods and techniques. In-depth training in specialized areas also is provided. The tutorial program has been designed to provide diversity and depth, and to appeal to researchers and practitioners.

Approach to conversion

The task in the case of the CHI tutorial collection consisted of gathering information from web pages on the Internet and generating the metadata from this information. A brief description of each tutorial presented at these conferences was found on the website for the conference [17]. The first step involved analyzing these web pages to determine what metadata could be derived from them and then converting it into XML records in the CITIDEL internal stream-of-records metadata format.

Since this was required to be an automated process, it also was essential to determine the common elements across different tutorial descriptions and years and to form an outline for the data that should be extracted from the web pages and where this data would fit within the target format, the CITIDEL internal stream-of-records metadata format.

The information contained on the web page for each tutorial of the conference had the following format: name of the tutorial followed by presentation and creator information. This then was followed by a brief description of the tutorial such as benefits, origins, features, audience, and instructor information.

Upon analysis of the web pages for the different tutorials across different years, it was determined that the best approach to this task was to generate a Perl script that would extract the information from the web page (essentially the metadata) and generate XML files which would contain this information in the CITIDEL internal stream-of-records metadata format. This Perl script would depend on the structure (HTML) layout of the page and require that this layout be consistent across the pages for different tutorials.

Difficulties encountered in conversion

As mentioned above, the script depended on the HTML code (structure) of the page remaining consistent across different tutorial pages. As is often the case with many web sites and web pages, this was not the case. Thus, completely depending upon parsing, based on the HTML structure of the page, was not feasible. Also, in the case of CHI 2000, the description of the tutorial was being served dynamically using cgi scripts. This further complicated the extraction process.

The major difficulty encountered was the layout of the metadata over the web pages. Each year, the conference had approximately 30 tutorials that were organized on the web site according to the day of presentation. Thus each of the years had different web pages corresponding to tutorials presented over different days of the conference. The structure of each tutorial consisted of a title followed by the names of the creators and other relevant metadata as described above.

This approach also required that the layout and structure as described above be consistent across different tutorial descriptions as well as different years. As it turned out, even though the structure followed a particular order across most of the tutorial descriptions, there were instances where the order was completely vague.

Overcoming these difficulties

The final source code contained a Perl script that extracts the metadata by parsing the metadata contained on the target web page of the tutorial descriptions. The approach followed consisted of a combination of parsing based on the HTML structure and keywords in the text of the description so as to extract the relevant metadata. Once these keywords/delimiters in the text were identified, the corresponding tags in the CITIDEL internal stream-of-records metadata format had to be determined.

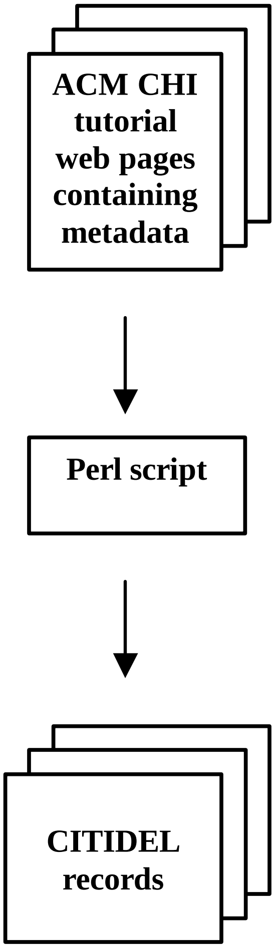

Figure 4: Conversion of ACM CHI tutorial metadata to CITIDEL records

The web page for each year was first parsed to generate collection level metadata for that particular year. This was accomplished by extraction of metadata based on the HTML structure of the web page. The next iteration was used to generate item level records for the different tutorials presented in that particular year. Here, extraction was based on the HTML structure as well as keywords in the text of the tutorial descriptions.

As mentioned in the previous section, in some cases, where the structure did not follow the layout as required, some minimal manual intervention was performed which involved addition of tags to complete the source HTML and make it well formed.

End Result

Parsing the web pages of the CHI tutorials helped in collecting both collection level as well as item level metadata for each of the years 2000, 2001, and 2002. The corresponding XML files containing XML records in the CITIDEL internal stream-of-records metadata format were generated for each of the tutorials presented at the conference.

Cross-links, which form an important component of the CITIDEL internal stream-of-records metadata format, between collection level records and the item level records, for each year, were generated. Person records for each of the creators of the tutorials also were generated.

2.4 HCI Bibliography: Human-Computer Interaction Resources

Description of the resource

Bibliographic resources can help students learn about HCI, researchers find relevant background, and developers find examples of the state of the art. It's easier to browse or search through a bibliography than to find publications in a library. It's a lot easier if the materials are online and Internet accessible.

Bibliographic information traditionally contains basic publication information about a published work: author(s), title, date, publisher, pages, etc. Often, bibliographic records include summary information such as keywords, abstracts, and section headings. In some cases, bibliographic records contain information about references - either their number or actual citations. Bibliographic records do not, however, contain the full text, tables, or figures in published works, so in terms of electronic publication, bibliographic records contain what is easy to represent about publications without addressing the more difficult issues of document content representation. Bibliographic records give you enough information to decide if a publication is worth getting for a full read.

There are two types of bibliographic collections: **collections on specific topics**, spanning many publications such as journals, conferences, books, etc., often annotated by the bibliographer, and **comprehensive collections**, covering many topics within a field.

The HCI Bibliography [18] is a free-access online bibliographic database on Human-Computer Interaction. The basic goal of the Project is to put an electronic bibliography for most of HCI on the screens of all researchers, developers, educators, and students in the field through the World-Wide Web and anonymous ftp access.

The HCI Bibliography is primarily a database of conference proceedings and journal volumes. Each module contains records of all the publications in the conference or journal volume. Sometimes a module is broken up to keep the file sizes manageable. (Some conferences have hundreds of papers, which would result in files with hundreds of thousands of characters, files that would create problems for many mailers.) Currently many conferences are covered back to about 1980, and many journals are covered back to their first volume. For example, the International Journal of Man-Machine Studies, recently renamed the International Journal of Human-Computer Studies, is covered back to 1969.

The HCI Bibliography also contains special files on books, edited collections of articles, reports, and videos on HCI. There are answers to frequently asked questions on publishers, professional organizations, and suggested readings. There is even an online copy of the ACM Computing Reviews Classification System of keywords.

The HCI Bibliography is on the World-Wide Web at: http://www.hcibib.org/

This is a free service. One does not need to pay any fees or be a member of any organization to use the HCI Bibliography, although some links to materials may require them.

As of December 2002, the HCI Bibliography had over 20,000 entries.

Approach to conversion

The HCI Bibliography is stored in a customized version of the UNIX refer format. So a book record looks like:

```
?? %T Title of the Book
?? %A First Author
?? %A Second Author
?? %D Date
?? %I Publisher
?? %C City
?? %P Number of Pages
?? %G ISBN
```

The entire HCIbib collection is contained in numerous .bib files. These files contain the raw data in the refer format explained above which is limited to single-character field descriptors. This means that some of the field descriptors are non-mnemonic, and that realistically, the data is limited to 26 field names, even though refer programs are case-sensitive.

The aim here was to build an OAI data provider for the HCIbib collection that then could be harvested by CITIDEL so as to obtain all the information.

Difficulties involved in conversion

As mentioned before, the primary aim was to build a data provider for the HCIbib collection. This required that the digital library/collection be either in the form of a database or in the form of XML files. Since the HCIbib collection was in neither form, this issue was most important.

Overcoming these difficulties

Since the closest equivalent to the two formats mentioned above was XML files, the first step taken was to convert the records contained in the .bib files to XML records in the Dublin Core format. This was accomplished using a Perl script.

The generated script results in a completely automated process. It used the Linux “wget” command to fetch a list of the available .bib files and compares this list with a local one. If there are any differences in them, which meant that new .bib files have been released, these individual files are then fetched and the records contained in them are converted to Dublin Core XML. These XML files then are placed in a directory over which the data provider runs.

The total number of files generated number 21900 which meant that the collection contained 21900 articles.

Once the metadata was obtained in the form of XML files in Dublin Core format, finding a data provider that could support such a large number of files was a challenge in itself.

Hussein’s OAI-PMH2 XMLFile File-based Data Provider was configured to run a data provider on top of these XML files. However, this did not turn out to be successful since the tool has a scalability limit of only 5000 files.

The next attempt was to try and implement the data provider using the OAICat tool [19] from OCLC. This is a multi-purpose data provider tool which has been successfully used in the past at Virginia Tech to implement data providers [20]. The new version of this tool contained a default file system implementation; this was adapted to run a data provider over the converted Dublin Core XML files.

Since the tool had been used before and is implemented in Java and thus keeps only a reference to the source files in memory when running, it was easy to customize. It adapted very well to the HCIbib collection in spite of the large number of XML files contained in it.

Initial testing for compliance was performed using the HTML interface of the tool. When the responses for the different verbs were customized and verified to be accurate, the next phase of testing was performed. This involved using Open Archives Initiative’s Repository Explorer [21] to test the data provider. Repository Explorer is a tool that is used to test data providers for compliance with the OAI-PMH version 2. It runs a series of validation tests to check whether a data provider has been successfully implemented.

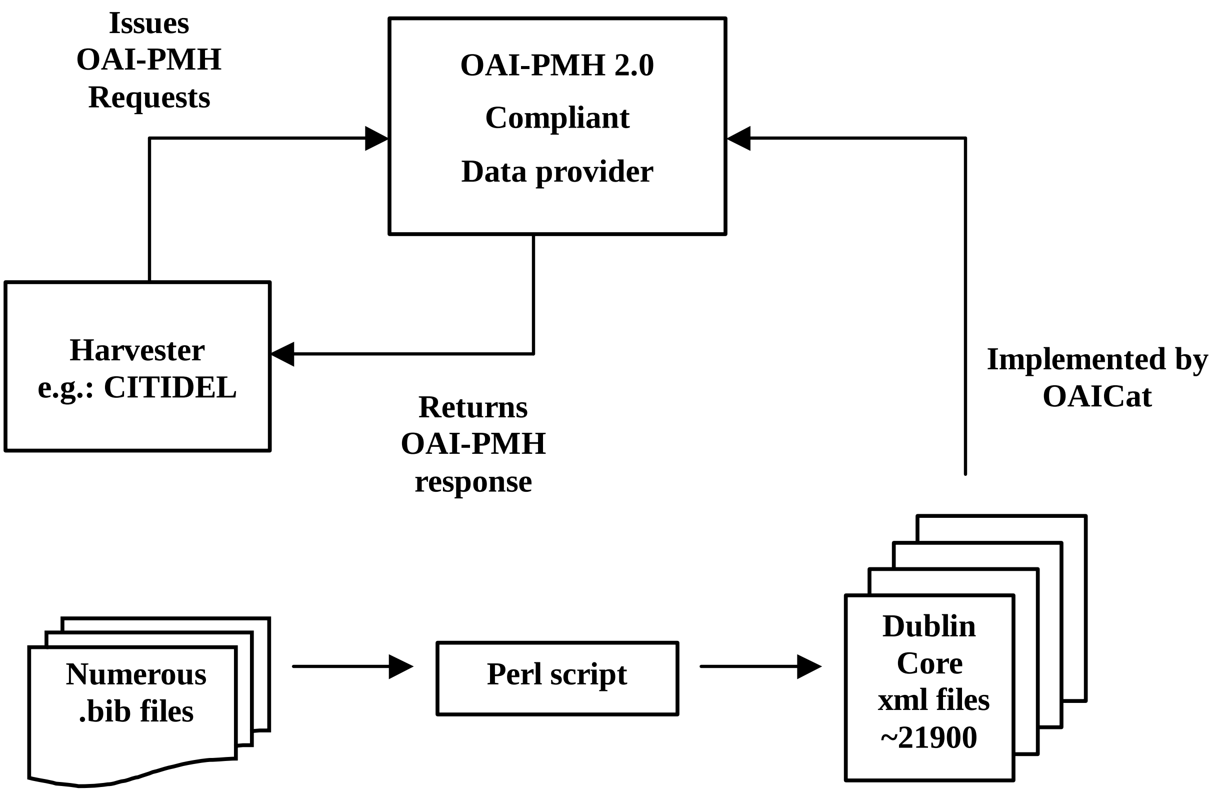

Figure 5: High-level architecture of the conversion of HCIbib data and implementation of a data provider over the XML files

End Result

?? The .bib files of the HCIbib collection were successfully converted to Dublin Core XML.

?? An OAI data provider was successfully implemented for the HCIbib collection. OAI's Repository Explorer verified this.

A sample record from the HCIbib collection and the corresponding record in the Dublin Core XML format is as shown below.

Original record in the UNIX refer format.

```
%M C.ASSETS.2000.1
%T Low Vision: The Role of Visual Acuity in the Efficiency of Cursor
Movement
%A Julie A. Jacko
%A Armando B. Barreto
```

```
%A Gottlieb J. Marmet
%A Josey Y. M. Chu
%A Holly S. Bautsch
%A Ingrid U. Scott
%A Robert H. Rosa
%B Fourth Annual ACM Conference on Assistive Technologies
%D 2000
%P 1-8
%I ACM
%K Low vision, Cursor movement, Icon size, Age-related macular
degeneration,
Search strategy, Graphical user interface, GUI
%* (c) Copyright 2000 ACM
%W      http://www.acm.org/pubs/articles/proceedings/assets/354324/p1-
jacko/p1-jacko.pdf
%X Graphical user interfaces are one of the more prevalent interface
types which exist today. The popularity of this interface type has
caused problems for users with poor vision. Because usage strategies of
low vision users differ from blind users, existing research focusing on
blind users is not sufficient in describing the techniques employed by
low vision users.
   The research presented here characterizes the interaction strategies
of a particular set of low vision users, those with Age-related Macular
Degeneration, using an analysis of cursor movement. The low vision
users have been grouped according to the severity of their vision loss
and then compared to fully sighted individuals, with respect to cursor
movement efficiency.
   Results revealed that as the size of the icons on the computer
screen increased, so did the performance of the fully sighted
participants as well as the participants with AMD.
```

The same record after conversion to the Dublin Core metadata format is as follows.

```
<?xml version="1.0" encoding="UTF-8"?>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                 http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:creator>Julie A. Jacko</dc:creator>
      <dc:creator>Armando B. Barreto</dc:creator>
      <dc:creator>Gottlieb J. Marmet</dc:creator>
      <dc:creator>Josey Y. M. Chu</dc:creator>
      <dc:creator>Holly S. Bautsch</dc:creator>
      <dc:creator>Ingrid U. Scott</dc:creator>
      <dc:creator>Robert H. Rosa</dc:creator>
      <dc:identifier>Fourth Annual ACM Conference on Assistive
      Technologies , 2000, Page(s) 1-8</dc:identifier>
      <dc:date>2000</dc:date>
      <dc:publisher>ACM</dc:publisher>
      <dc:subject>Low vision, Cursor movement, Icon size, Age-related
      macular degeneration, Search strategy, Graphical user interface,
      GUI </dc:subject>
      <dc:identifier>C.ASSETS.2000.1</dc:identifier>
      <dc:source>hcibib</dc:source>
```

```
        <dc:title>Low Vision: The Role of Visual Acuity in the Efficiency
        of Cursor Movement </dc:title>
        <dc:rights>(c) Copyright 2000 ACM</dc:rights>
        <dc:identifier>http://www.acm.org/pubs/articles/proceedings/asset
        s/354324/p1-jacko/p1-jacko.pdf</dc:identifier>
        <dc:description>Graphical  user  interfaces  are  one  of  the  more
        prevalent  interface  types  which  exist  today.  The  popularity  of
        this  interface  type  has  caused  problems  for  users  with  poor
        vision. Because usage strategies of low vision users differ from
        blind  users,  existing  research  focusing  on  blind  users  is  not
        sufficient  in  describing  the  techniques  employed  bylow  vision
        users. The research presented here characterizes the interaction
        strategies  of  a  particular  set  of  low  vision  users,  those  with
        Age-related  Macular  Degeneration,  using  an  analysis  of  cursor
        movement. The low vision users have been grouped according to the
        severity of their vision loss and then compared to fully sighted
        individuals, with respect to cursor movement efficiency. Results
        revealed  that  as  the  size  of  the  icons  on  the  computer  screen
        increased,   so   did   the   performance   of   the   fully   sighted
        participants    as    well    as    the    participants    with    AMD.
        </dc:description>
</oai_dc:dc>
```

3. Conclusion

3.1 Results

The goal of this study was to assist CITIDEL by increasing the number of collections available to it. This study focused on four such collections:

1. Digital Bibliography and Library Project (DBLP)
2. ACM SIGCHI S.P.A.C.E. Student Exhibition
3. ACM SIGCHI CHI Tutorial Presentations
4. HCI Bibliography: Human-Computer Interaction Resources

Each of these collections was worked on as described in earlier sections. An approach was decided on, which was then followed and difficulties were overcome as described in the sections related to the collections.

Thus, these collections were successfully converted such that CITIDEL should derive significant benefit. As a direct result of this, most of the metadata was converted directly into the CITIDEL internal stream-of-records metadata format such that it can be directly imported by CITIDEL.

The particular collection and the number of records contained in each that were successfully converted for use with CITIDEL are summarized in a table as shown below.

3.2 Table

Collection	**No of Records**
DBLP	320009
ACM S.P.A.C.E Student Exhibition	320
ACM CHI Tutorials	94
HCIbib	21900

3.3 Tools used

Tool	Description
org.jdom package	There is no compelling reason for a Java API to manipulate XML to be complex, tricky, unintuitive, or troublesome. JDOM is both Java-centric and Java-optimized. It behaves like Java, it uses Java collections, it is a completely natural API for current Java developers, and it provides a low-cost entry point for using XML. While JDOM interoperates well with existing standards such as the Simple API for XML (SAX) and the Document Object Model (DOM), it is not an abstraction layer or enhancement to those APIs. Rather, it seeks to provide a robust, lightweight means of reading and writing XML data without the complex and memory-consumptive options that current API offerings provide.
TableExtract	HTML::TableExtract is a subclass of HTML::Parser that serves to extract the textual information from tables of interest contained within an HTML document. The text from each extracted table is stored in table state objects that hold the information as an array of arrays that represent the rows and cells of that table.
OAICat	The OAICat Open Source project is a Java Servlet web application providing an OAI-PMH v2.0 repository framework. This framework can be customized to work with arbitrary data repositories by implementing some Java interfaces. Demonstration implementations of these interfaces are included in the webapp distribution.

3.4 Lessons learnt

Web pages, including those belonging to the same web site, often do not follow a consistent layout or structure for their content. This was observed with the S.P.A.C.E. and the ACM SIGCHI collections, where in some cases, structure was not followed at all.

The above point is also related to patterns of data. When dealing with different collections, patterns are often observed in the layout of the data. These patterns are important because extraction rules can often be based on patterns. However, in reality, it is not simple to establish patterns because of inconsistencies that are omnipresent.

Converting metadata from one form to another, or trying to extract it, cannot always be done transparently. A small amount of manual intervention is sometimes required. This is especially true with collections existing in “loosely-typed” formats such as on web pages. Due to the “loosely-typed” nature of HTML, inconsistencies in the source HTML had to be taken into consideration, while extracting metadata, with both the S.P.A.C.E. and ACM SIGCHI collections.

One of the most important lessons learnt from this project deals with the issue of data quality. There are many different and unique collections that exist, but what needs to be determined is the quality of the data they deal with. One question that often needs to be asked when dealing with a particular collection is: “How good is the quality of the data / metadata and how helpful will it be after conversion to the target format?”

When dealing with continually updated collections, to derive maximum benefit from the data / metadata, it is advantageous to find a solution that takes into account the change in the data with time. Such a solution might require holding some state information regarding old data that has already been dealt with. In the case of the HCIbib collection, new .bib files are released and added to the collection at regular intervals. The Perl script generated to deal with this collection performs a comparison between the current list on the website and a locally maintained list that keeps track of old data that has already been dealt with. New Dublin Core XML records are generated only for records contained in newly released .bib files.

When dealing with static collections (e.g., data regarding conferences, exhibitions), a one-time conversion of data to the target format is sufficient as opposed to the above point. This was observed with both the S.P.A.C.E. and ACM SIGCHI tutorial collections. Since these were static collections that would not change with time, a one-time conversion to the CITIDEL internal stream-of-records metadata format was performed.

3.5 Open problems

Our one open problem is that there are missing slide sets for the years 1998 and 1999 of the S.P.A.C.E. collection, since information about the slides from these years was missing from the web site itself.

4. Acknowledgements

NSF grant number for CITIDEL: DUE 0121679

I would like to thank the following people who have played an important part in the completion of this project:

?? Dr. Edward A. Fox for his continued support and guidance before the beginning of as well as throughout the duration of the project.

?? Aaron Krowne, Rohit Kelapure, and many others in the Digital Library Research Laboratory (DLRL) who have been patient enough to guide me and answer every possible doubt that I have had.

?? Jeff Young from OCLC who was kind enough to answer my doubts regarding the process of customizing the OAICat tool for my particular task.

5. References:

1. All scripts created and used for this study can be found at:
   http://crawler.cc.vt.edu/dblp/scripts

2. Computing and Information Technology Interactive Digital Educational Library (CITIDEL):
   http://www.citidel.org/

3. National Science Digital Library (NSDL):
   http://www.nsdl.org

4. Open Archives Initiative (OAI):
   http://www.openarchives.org/

5. CITIDEL internal stream-of-records metadata format:
   http://tennessee.cc.vt.edu/~akrowne/citidel/schema/

6. Dublin Core:
   http://www.dublincore.org

7. IMS Global Learning Consortium, Inc (IMS) metadata:
   http://www.imsglobal.org/metadata

8. IEEE Learning Technology Standards Committee (LTSC) Learning Object Model (LOM):
   http://ltsc.ieee.org/

9. DBLP website:
   http://www.informatik.uni-trier.de/~ley/db

10. Mirror package:
    http://sunsite.org.uk/packages/mirror/

11. JDOM:
    http://www.jdom.org/

12. ACM SIGGRAPH:
    http://www.siggraph.org/

13. ACM SIGGRAPH S.P.A.C.E. website:
    http://www.siggraph.org/education/space/space.htm

14. CPAN website:
    www.cpan.org

15. ACM website:
    http://www.acm.org/

16. ACM SIGCHI website:
    http://www.acm.org/sigchi/

17. ACM CHI Tutorials:
    http://sigchi.org/education/tutorials2go.html

18. HCI Bibliography:
    http://www.hcibib.org

19. OAICat:
    http://www.oclc.org/research/software/oai/cat/shtm

20. Project report for the American South 1 A project (fall 2002):
    http://rocky.dlib.vt.edu/~cs5604/fall_2002/AmericanSouth1A/

21. OAI Repository Explorer:
    http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai