Developing an XML framework for an electronic document delivery system

Shien-Chiang Yu and
Ruey-Shun Chen

The authors

Shien-Chiang Yu and **Ruey-Shun Chen** are both based at the Institute of Information Management, National Chiao-Tung University, Taiwan.

Keywords

Interfaces, Document supply, Information services, Internet, EDI, Information retrieval

Abstract

The Internet has forced libraries to consider how to assist users to rapidly retrieve information. Such a consideration has accelerated the development of electronic publishing and has positioned the library as mediator between users and providers: archiving information circulation and providing secure copyright clearance through an efficient electronic document delivery and payment mechanism. This work develops an Extensible Markup Language (XML) framework for electronic document delivery that offers a novel electronic document delivery system and also locates publishers who can provide the copyrighted material in an electronic format via the OPAC. The proposed electronic document delivery system has four functions: (1) it enables the electronic document payment; (2) it shortens the time between inquiry and electronic document retrieval; (3) it anticipates the changing role of libraries; and (4) it reduces the printed collection load of libraries.

Electronic access

The research register for this journal is available at **http://www.mcbup.com/research_registers**

The current issue and full text archive of this journal is available at **http://www.emerald-library.com/ft**

The Electronic Library
Volume 19 · Number 2 · 2001 · pp. 102–110
© MCB University Press · ISSN 0264-0473

Introduction

Library users previously had to search through the individual library catalogs or through the union catalog of bibliographical centers to find what they wanted. However, the Z39.50 Information Retrieval Protocol (Hakala, 1996) has provided for automation systems and bibliographic databases. Z39.50 has enormously promoted the scope of information retrieval, since it allows a PC or terminal with access to a local computer to search the database of a remote computer according to the procedures and formats of the local system. Thus, traditional library catalogs provide bibliographic services rather than full-text services and they cannot compete with the speed and convenience of the computerized system.

Traditional libraries acquire printed media via their document management and collection departments. These libraries face space and financial restrictions, since the amount of holdings rapidly expand with the costs of individual publications while library budgets are continually decreased (Luther, 1999, pp. 179-181). Thus, libraries must efficiently collect information and utilize the Inter Library Loan (ILL) service to offset their limitations. However, it is often too time consuming to be of practical advantage.

Electronic publication is gaining in popularity because it makes the library collection policy become "access" to replace "ownership". Indeed, this reminds the library that the user willing to obtain the information is more essential to the libraries holding the information. A library can save money and space by efficiently utilizing electronic publications, as document providers (publishers, vendors, etc.) rent their electronic media instead of selling them. The number of electronic documents has increased so dramatically that document delivery and Information Retrieval (IR) services are continuously being updated to access information more quickly.

Extensible Markup Language (XML) offers a unique combination of flexibility, simplicity, and readability by both humans and machines. XML uses a reasonably concise syntax that provides developers with

The authors would like to thank the National Science Council, Taiwan, R.O.C., for financially supporting this research under Contract No. NSC89-2416-H009-012.

enormous power. XML has clear structures that make it useful for larger projects as it can be massaged, manipulated, processed, fragmented, and rebuilt far more easily than previous formats.

This study develops an XML framework for an electronic document delivery (EDD) system by enhancing existing library automatic systems with Library Online Public Access Catalog (OPAC) facilities. Users can retrieve bibliographic information through the library mediation service and they can directly transmit their needs into the document publisher or owner's system. The users can obtain electronic full-text information to recall something in the system, whether it is free or not. Libraries can also provide online services, even if they do not have the desired texts. Such a system greatly facilitates the digital library.

The electronic document publisher makes online electronic exchanges through the library online catalog. Thus, libraries can reduce the burden of the collection and also charge for referral services. Libraries can utilize this fee to broaden their services while the electronic document publisher benefits from the user downloading the document. These benefits encourage the publisher to more quickly issue the electronic version of the document.

Relevant literature

This section explores relevant literature on both the XML and the EDD system.

XML

The Internet community is increasingly focusing on XML as its next-generation data representation vehicle XML and Resource Description Format (RDF): the promise and the reality of new Web architectures. XML supports language- and platform-neutral facilities and defines a data format for structured document interchange on the Web. XML is a simple subset of the Standard Generalized Markup Language (SGML) derived by the World Wide Web Consortium (W3C) in February 1998. Unlike HTML, which defines a fixed set of tags, XML enables customized markup languages to be defined with application-specific tags that represent information in such application domains as chemistry, electronics, and general business. Using such tags to delimit individual data items or groups of data called elements permits the programs to easily identify and process the data within a XML document (W3C, 1998).

Both libraries and document providers can easily modify XML-exchange information in existing applications, since XML supports the definition of language-neutral and platform-neutral facilities. Our electronic document delivery system model prefers XML for two reasons: language theory and practical application. XML enables developers to create and manipulate their own tags and it also works smoothly with Cascading Style Sheets (CSS) to enable developers to present information as it is originally structured. XML is an excellent format for interchanging data, since browsers (like IE5.0) can read XML data. XML's ubiquity is more practically commercial as the Web has made everyone appreciate the power of markup languages, practically ensuring the widespread adoption of XML as HTML's heir apparent.

EDD system

The notion behind EDD dates back to the first computers. Cawkell (1991, pp. 34-7) reported that:

> The phrase "electronic document delivery system" self-evidently implies the supply and reproduction electronically of the kind of information usually provided in the form of print on paper.

Three generations of EDD systems can currently be distinguished: systems based on online ordering, non-integrated supply-driven image-based systems, and integrated stand-alone image-based systems (Roes and Dijkstra, 1994, pp. 13-14):

(1) The first generation – online ordering, such as DIALOG and ESA/IRS, offer this kind of service. An application connected to the reference databases could produce a work list for document delivery and operations. Final delivery is by ordinary mail or facsimile to the applicants. A vital disadvantage of this system is that it inefficiently processes requests. Each time an article is requested, the library staff must go to the shelves, locate the article, and make photocopies, even if it is a request for an article previously requested.

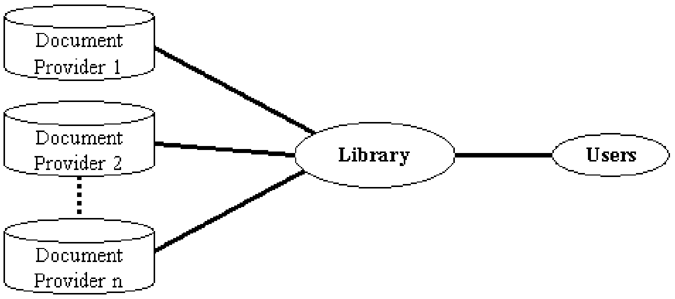
(2) The second generation, supply-driven image-based systems, introduces a supply-driven approach to storing articles as images. The ADONIS system, for example, advocates the scanning and storage of articles that are expected to be in frequent demand. Nevertheless, real delivery still was the disadvantage.
(3) The third generation, scalable stand-alone systems, are still based on images, but a demand-driven approach is applied to reduce overhead. The delivery service is expanded at the end-users PC so it can display images of the referenced article when a user finds a reference. These images can be presented or downloaded to the PC by clicking a button. Some examples of an approach with third-generation characteristics are CARL's Uncover, OCLC ContentFirst, ArticleFirst, ContentsAlert, Faxon Finder, SWETSCAN, EBSCOdoc, ISI's Current Contents.

Unfortunately, image-based systems are not open to future developments and the ILL services and existing document delivery procedures are poorly integrated. Users must individually query databases and then make decisions of whether to make delivery or download materials.

The simplest method to resolve these problems is to integrate the table of contents of every document delivery producer. Front-end users can retrieve all databases at one search and the best intermediary of the organization is the library (Figure 1).

Most current EDD systems prefer to implement electronic journals and "from paper to CD-ROM documentation" rather than completely relying on the Web. A new EDD system model is proposed to resolve the drawbacks of current EDD systems.

Figure 1 The library should be the intermediary of information between users and document providers



The proposed EDD system

Critical EDD system technologies are not yet adequately developed and most publishers still publish printed materials more than any other material. There are two basic reasons for this delay (Berghel, 1999, p. 19):
(1) The primary intended venue for electronic publishing, the Web, lacked secure HTTP transactions until 1995. Without secure transactions, selling via online would entail excessive risk due to network hijacking, such as digital eavesdroppers, packet sniffers, and other network nematodes.
(2) Nobody knew how to develop a sound business plan for electronic publishing.

Therefore, EDD systems must consider copyright, payment and how to display the contents of the different media on the front-end users. This concept places more emphasis on quickly delivering the full-text to the user than the format of the electronic information and the hardware. Users will decide which document to download when the browser displays the entire results of the database inquiry. With a secured commercial environment and developments related with online ordering, electronic publishing will make a huge progress.

The EDD system XML framework design

The present system must be simplified to reduce the impact of implementing the EDD system and to minimize the total cost. XML was adopted for data exchange, transmission, processing, and presentation:

- The library system employed XML/EDI (Electronic Data Interchange) to communicate with document providers.
- The Web Server (OPAC) downloaded XML data to the client browser. The XML metadata detailed how to access the document provider.

Consequently, the EDD system model offered an architecture and system to address the three major issues in document searches using XML framework:

- Transferring the result-set of a query into XML format.
- The metadata about the result-set.
- The user can link into the system of document-provider directly, through the XML document which downloads from the OPAC.

The EDD system architecture

The system uses the browser as a client's interface to communicate with a HTTP Web server. On the server side, an OPAC Web server joins with the existing library system to offer the EDD system front-end search function. The EDD system must use an advanced set of computational and network tools from current library systems as illustrated in Figure 2. An exchange of information between document providers and users (document consumer) via an XML/EDI agent is one such front-end query function (McCauley-Bell, 1990, pp. 135-47).

The agent on the library side must receive XML/EDI data that is transferred from the document provider agent. The XML/EDI data includes:

- A new article lists of providers or systems that can automatically convert the data into a database.
- A service log: the commercial-transaction between the user and the document provider.

The agent on the document provider side must:

- control access;
- process the requested document;
- plug-in the dispatch;
- encrypt the document and generate secret key;
- control payment.

Operation of the proposed EDD system

The cataloging record adds publishers' information including the defaults such as URL, document number, encrypted algorithm, type of material, kind of payment and price. Similar to the copies' management in the library automation system, users may also see if some document providers provide the same article in various languages, material-types, providers, etc., via OPAC.

The library-site will download the bibliographic XML-formatted data that he/she requested and the browser will display the search result on the screen when a user connects to the library system OPAC to query a bibliographic/journal article (see Figure 3). The user can, in accordance with his or her needs and the type of library service, decide on a method of acquiring the article: booking or loan from a library, directly displaying the content if the article is saved in the library's Web page or database.

The user may use the program they plugged in to download the article from the library automation system or the document provider if the article is stored in an electronic medium through the document providers (or publisher). The security and payment of a downloaded document may be followed by electronic commerce – except for free articles.

The browser may control related browsing programs and copyright by plug-in, for instance, presenting from illegal copy, rebuild

Figure 2 Integrated EDD system architecture

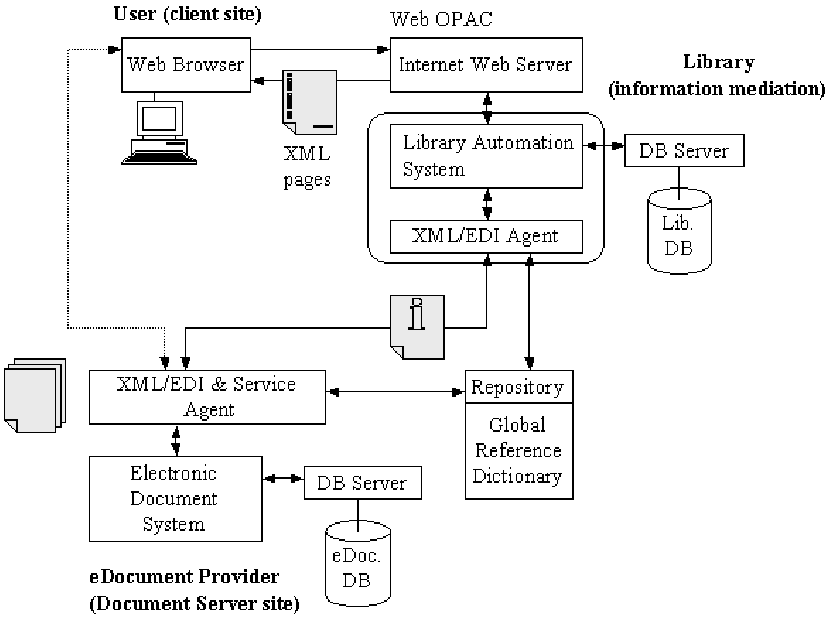
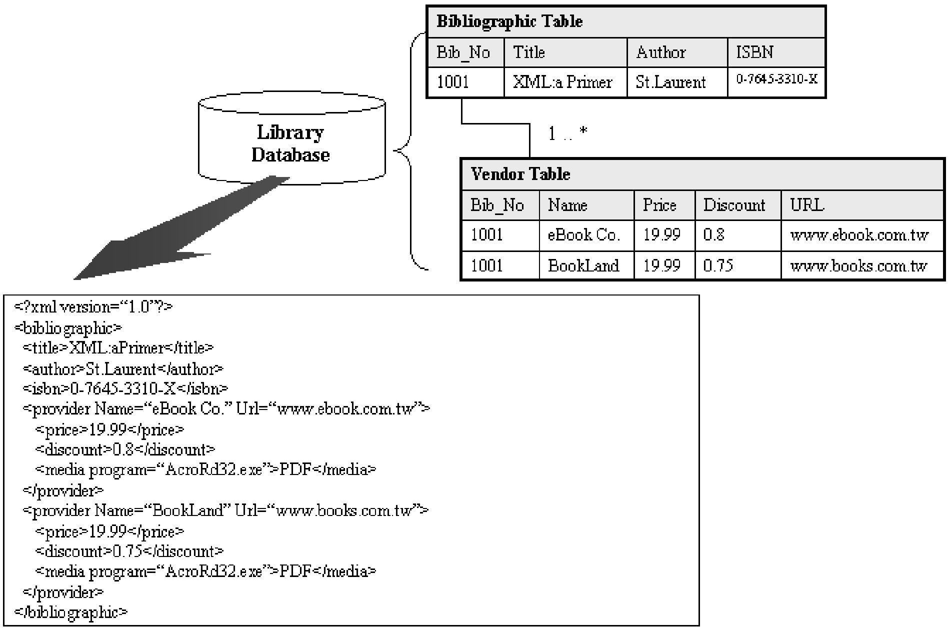


Figure 3 OPAC transfers the search result to the client via the XML format



Bibliographic Table			
Bib_No	Title	Author	ISBN
1001	XML:a Primer	St.Laurent	0-7645-3310-X

Vendor Table				
Bib_No	Name	Price	Discount	URL
1001	eBook Co.	19.99	0.8	www.ebook.com.tw
1001	BookLand	19.99	0.75	www.books.com.tw

```
<?xml version="1.0"?>
<bibliographic>
  <title>XML:aPrimer</title>
  <author>St.Laurent</author>
  <isbn>0-7645-3310-X</isbn>
  <provider Name="eBook Co." Url="www.ebook.com.tw">
    <price>19.99</price>
    <discount>0.8</discount>
    <media program="AcroRd32.exe">PDF</media>
  </provider>
  <provider Name="BookLand" Url="www.books.com.tw">
    <price>19.99</price>
    <discount>0.75</discount>
    <media program="AcroRd32.exe">PDF</media>
  </provider>
</bibliographic>
```

(screen capture, format transfer), and time limitation.

The elements of the proposed EDD system

The proposed EDD system model integrates users and library automation systems with the existing system of most electronic document providers. It involves four interconnected elements:

(1) Data exchange between the browser, library system, and provider systems.
(2) Copyright protection.
(3) A payment mechanism.
(4) The browser's plug-in.

Data exchange

The automation system, either library or electronic document provider, must add a Web-based EDI product in order to communicate with each other. XML/EDI (Webber, 1998, p. 39) is the best approach to match this necessity.

XML/EDI provides the infrastructure of a wide variety of systems from searchable online catalogs to robust machine-to-machine transaction subsystems. The system can create a Document Type Definition (DTD) that formally defines the structure of EDI messages without having to use it to validate well-formed XML documents. The system can simply place EDIFACT/X12 messages in an XML shell element or parts of such an element.

The main difference between XML/EDI and other methods is that an XML/EDI system can encode the information involving a document significantly more precisely than was previously possible in earlier formats. XML/EDI transactions are self-describing applications that employ XML elements and define DTD to process XML/EDI documents so it can "understand" a transaction only by accessing the content of the transaction.

XML/EDI seems well placed to be the technology of the future for the following reasons (Peat and Webber, 1997):

- It has open standards.
- It provides for self-describing transactions (XML).
- It allows tool vendors to improve existing products.
- It readily interfaces with legacy systems.
- The framework uses an evolving best-of-breed philosophy – i.e. dynamic shared dictionaries.
- It allows for object-based documents – where the data and rules reside together.
- It is cheaper and easier to implement.

Access to "interactive" transactions is enabled by the Web rather than "system" or "batch" transactions.

Payment

The types of payment for receiving electronic documents include:

- Free of charge: some academic papers, partial dissertations, or electronic documents on Web pages are free.
- Payment by the library: academic libraries provide document service to faculty and students. The fees are generally paid by the library itself or directly by ordered online databases (ex. ABI/INFORM, IDEL, EBSCO), query or downloaded documents.
- Payment by the user: this is similar to business-to-consumer electronic commerce.

Free articles are generally in a library's electronic media. The simplest method to access resources on another Web site is to create a Resource Description Format (RDF) through the resource description link to that document.

The payment mechanism must calculate the amount of usage when payment is required to access a document from the provider (publisher or vendor). The provider must incorporate a process to calculate the actual amount of usage between providers and the library.

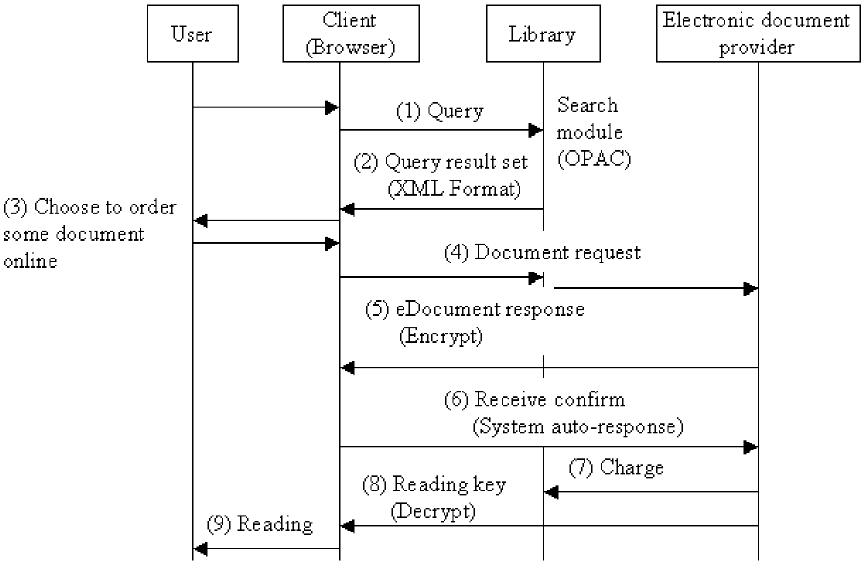
The three different types of charges are illustrated in the sequence diagrams shown in Figure 4.

Free of charge. We do not discuss this situation, because these systems do not involve security and payment.

Payment by the library:

(1) The user connects to the library OPAC system, and query data through the browser.
(2) In accordance with the request from the browser, the OPAC system searches the database and acquires the result of the query. Before responding to the browser, the OPAC system must recompose these data into XML format.
(3) After the browser receives the XML document in response from the OPAC system, the XML document encompasses all of the information on the result of the query (Figure 3). The user can determine which data she/he wants.
(4) When the user moves the mouse to click the item, which displays on the browser, the browser will link into the original data source (Electronic document provider (EDP)) and pass some information. The EDP system will transmit the original data source (document source) to the browser, if the library has a contract with the EDP.
(5) For security over the Internet, the EDP system must encrypt the document.
(6) When the browser receives the document, it must respond with a

Figure 4 EDD system's sequence library payment diagram


User
Client (Browser)
Library
Electronic document provider
(1) Query
Search module (OPAC)
(2) Query result set (XML Format)
(3) Choose to order some document online
(4) Document request
(5) eDocument response (Encrypt)
(6) Receive confirm (System auto-response)
(7) Charge
(8) Reading key (Decrypt)
(9) Reading

confirmation message to the EDP system.

(7) The EDP system, according to the confirm message will record the fee, and charge the library the fee.

(8) The EDP system transmits the key of decryption to the browser.

(9) The browser decrypts the document and displays content.

Payment by the user. Steps from (1) to (4), and from (7) to (10) are as same as those in Figure 4. (5) In order to be a Trusted Third Party (TTP), the OPAC system records the transaction log. The user can attain non-repudiation control via public key encipherment. (6) The EDP system can achieve the payment by the access control via only registered users being able to access the EDP system. When users login to the EDP system and download the original data source (document source), the system will record the fee, and charge the user for the fee sometime later. (11) In view of the Figure 5 sequential diagram, the library is the intermediary between users and EDPs. The library may charge EDPs for the middleman fee.

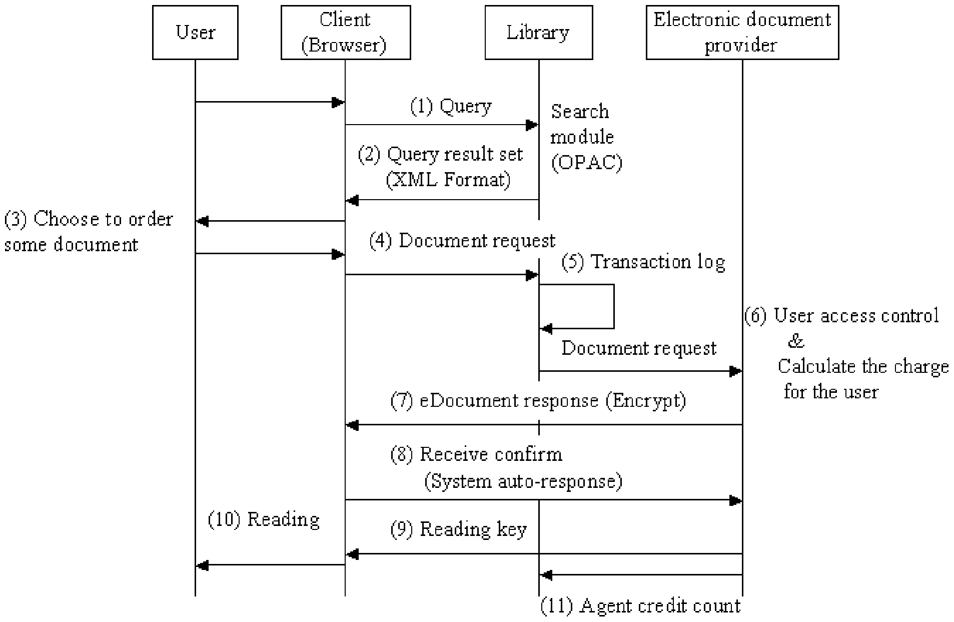
Copyright

Electronic documents must include a copyright protection mechanism when they are in transmission to prevent the user from duplicating or converting the document without authorization. The major methods of copyright protection include cryptography and digital watermarking:

- *Cryptography.* This service ensures a document is available only to those authorized through public key, secret key, or security token crypto methods. Electronic documents are enciphered by a provider's secret key and randomly generated for the user as illustrated in Figures 4 and 5. The plug-in program cannot decipher and display, unless the client has received the secret decipher key. The user can attain non-repudiation control via public key encipherment or through the library to be the TTP by paying the fee.

  A digital signature is a cryptographic check-value that protects against forgery by the recipient and is computed using a private key and an asymmetric cryptographic algorithm (ISO/IEC 10181-1, 1996, p. 11). Digital signatures are employed for proof of origin and maintaining integrity. The document provider (the origin) could attach time-notation before computing the digital signature so the digital signature could correctly ensure the document's integrity. Digital signatures also include a non-repudiation function so the user (recipient) can verify that the sender is the real originator of the product.

Figure 5 EDD system's sequence user payment diagram


User
Client (Browser)
Library
Electronic document provider
(1) Query
Search module (OPAC)
(2) Query result set (XML Format)
(3) Choose to order some document
(4) Document request
(5) Transaction log
Document request
(6) User access control & Calculate the charge for the user
(7) eDocument response (Encrypt)
(8) Receive confirm (System auto-response)
(10) Reading
(9) Reading key
(11) Agent credit count

- *Digital watermarking technology*. Digital watermarking technology is the simplest way to prove the origin of an electronic document (visible and invisible). Digital watermarking technology can add a mark, signature, or any graphic hidden into the electronic media. The digital watermarking remains intact even if someone revises the image or document after downloading. Therefore, a notary could determine ownership in an intellectual property dispute.

Plug-in

Cascading Style Sheets (CSS) and Extensible Style Language (XSL) can combine with XML data that is transferred from OPACs to enable the browser to properly display its inquiry results. The browser requires the necessary plug-in to read the electronic documents because they have many distinct formats (e.g. PDF, images), various crypto-methods (e.g. digital watermarking, secret key, public key), reading limitations (e.g. times, user), and payment types (e.g. member, credit card). The document provider must support the plug-in if an electronic document has special requirements. Plug-ins can also reduce the provider's system modification. The plug-ins do not have to modify an electronic document provider's system very much. The plug-ins will simplify a complex design to satisfy the multi-tier architecture if they are part of a new system.

Comparison with other systems

The framework of the modified EDD system can help the library improve the traditional EDD system services. Some comparisons of our results with other EDD systems, e.g. the NAILDD Project (Barrett and Jackson, 1993) and the ARIADNE system (Roes and Dijkstra, 1994) are provided in Table I.

Additional work is being pursued in the following directions:

- An analysis of the system approach for the more complex electronic publisher.
- An agent-based technique between library and document provider.
- An implementation of the system approach in a prototype system.

Conclusion

Information-retrieval systems must be able to handle increasing volumes of articles and source materials. The XML markup technique employed to develop the proposed modified EDD system offers an integrated document query and retrieval service for end-users, electronic document payment options, and it promotes the recall-rate, shortening the time between inquiry and retrieval of the electronic document. This model repositions the information intermediary and reduces the collection load. This system can also completely

Table I A comparison of our results with other EDD systems

System functions	XML framework EDD system	Other EDD systems
Flexibility	The system can extend the service region by connecting two library or document provider systems	The system can connect to other libraries or document provider systems but only when these systems are very similar
Convenience	The user can acquire all of the information services through only one library Web query interface	The user must login and individually search each database. One cannot directly probe all of the linked document resources
System Outlay	The library system does not need substantial modifications, but it must extend some functions under the purpose of agent	Every module is an absolute work and there are no extensions from an existing library's automation system
Technology	XML and Agent technology are combined to conform to the applied IT trend	General network information service technology
Loading	The library does not need extra employees as it easily achieves full-automation. A library must have someone to manually key in data if the electronic document provider operates without automatically exporting new cataloged data by XML/EDI	A library must maintain and update any databases it purchased. However, a library does not need to do this if it is simply the intermediary between users and electronic document providers

combine an EDD system with a library's automatic system to archive information circulation and provide secure copyright fee through an efficient EDD and payment method. Thus, this system should also accelerate the development of electronic publishing.

References

Barrett, G.J. and Jackson, M.E. (1993), "Access and technology program/NAILDD project", *Association of Research Libraries*, Available at http://www.arl.org/access/naildd/naildd.shtml

Berghel, H. (1999), "Value-added publishing", *Communications of the ACM*, Vol. 42 No. 1, p. 19.

Cawkell, A.E. (1991), "Electronic Document Supply Systems", *Journal of Documentation*, Vol. 47 No. 1, pp. 34-7.

Hakala, J. (1996), "Z39.50-1995 Information retrieval protocol: an introduction to the standard and its usage", Available at http://renki.helsinki.fi/z3950/z3950pr.html

ISO (1996), ISO/IEC 10181-1 *Information technology – Open Systems Interconnection – Security Frameworks for Open Systems: Overview*.

Luther, J. (1999), "Electronic book '98 – turning a new page in knowledge management: NIST conference", *Library Collections, Acquisitions, & Technical Services*, Vol. 23 No. 2, pp. 179-81.

McCauley-Bell, P. (1990), "Intelligent agent characterization and uncertainty management with fuzzy set theory: a tool to support early supplier integration", *Journal of Intelligent Manufacturing*, Vol. 10, pp. 135-47.

Peat, B. and Webber, D. (1997), "Introducing XML/EDI: the e-Business framework". Available at http://www.geocities.com/WallStreet/Floor/5815/start.htm

Roes, H. and Dijkstra, J. (1994), "Ariadne: the next generation of electronic document delivery systems", *The Electronic Library*, Vol. 12 No. 1, pp. 13-20.

W3C (1998), *Extensible Markup Language (XML) 1.0: W3C Recommendation*, Available at http://www.w3.org/TR/1998/REC-xml-19980210

Webber, D.R. (1998), "Introducing XML/EDI Frameworks", *EM – Electronic Transactions, Electronic Markets*, Vol. 8 No. 1, pp. 38-41.