

ISSN 1340-7287

Digital Libraries, No.12

Proceedings of
International Joint Workshop
on
Digital Libraries 1998

7-9 September 1998
Asian Institute of Technology, Thailand

デジタル図書館, No.12

ISSN 1340-7287

Digital Libraries, No.12

Proceedings of
International Joint Workshop
on
Digital Libraries 1998

7-9 September 1998
Asian Institute of Technology, Thailand

デジタル図書館, No.12

Organizers:

- Asian Institute of Technology, Thailand
- Faculty of Engineering, Kasetsart University, Thailand
- The Thai Library Association Under the Royal Patronage of
H R H Princess Maha Chakri Sirindhorn, Thailand

Sponsors:

- Asian Institute of Technology, Thailand
- Faculty of Engineering, Kasetsart University, Thailand
- The Thai Library Association Under the Royal Patronage of
H R H Princess Maha Chakri Sirindhorn, Thailand
- University of Library and Information Science, Japan
- National Center for Science Information Systems, Japan

Digital Libraries, No.12, September 1998

Thomas Baker

Asanee Kawtrakul

Vilas Wuwongse

Asian Institute of Technology, Bangkok, Thailand

ISSN 1340-7287

<http://www.DL.ulis.ac.jp/DLjournal/>

<http://beethoven.cpe.ku.ac.th/ijwd198/DLjournal/>

Message from Organizing Co-Chairs

The International Joint Workshop on Digital Libraries 1998 (IJWDL'98) is composed of School on Digital Libraries and the Twelfth Workshop on Digital Libraries, a series of workshops that until now have been held at the University of Library and Information Science in Tsukuba, Japan. IJWDL'98 aims at providing an international forum for researchers and developers together with librarians and practitioners, particularly those in Thailand and this region, to share their experiences on trends in Digital Libraries (DLs). It is the first international meeting on DLs ever held in Thailand.

The Thai Library Association, Kasetsart University and the Asian Institute of Technology have cooperated as hosts of this workshop. DLs use advanced information technologies to organize information and make it accessible to users. They will be an important element of any National Information Infrastructure. This meeting will therefore be beneficial for Thailand and should raise the interest of Thai colleagues in DL research and development.

Our workshop would not have been a success without the kind assistance of Profs. Koichi Tabata and Shigeo Sugimoto, the kind contribution of all tutorial lecturers and the hard work of the Program Committee, co-chaired by Drs. Thomas Baker and Asanee Kawtrakul. We also wish to express our deep appreciation to the sponsors for their generous support.

The green surroundings on the campus of the Asian Institute of Technology should combine with the program to make this meeting unforgettable.

Chutima Sacchanand
Vilas Wuwongse
Organizing Co-Chairs

Message From the Program Committee

IJWDL'98 is the twelfth in a series of workshops that have been held since 1994 at the University of Library and Information Science in Tsukuba Science City. As the first to be held outside of Japan, this workshop was designed for a somewhat broader audience than its predecessors. We have sought to combine a traditional scientific workshop, with its focus on new ideas and applications, with a more general School of Digital Libraries for interested professionals in Southeast Asia who may be approaching the subject for the first time.

The four tutorials, accordingly, provide a "road map" to underlying developments and trends. Starting literally with nuts and bolts, Nobuhito Yamamoto introduces the basic technologies of computer and information networks. The technologies of digital libraries, of course, continue to evolve from year to year, even month to month. Kikuta Masahiro and Youichi Shibata focuses on how the multimedia document and information formats of the World-Wide Web, young as it is, are already being "reinvented" with sophisticated new functionality for advanced scholarship and digital commerce. Indeed, library science itself is being largely reinvented to handle world-wide repositories of information objects in an astounding variety of digital and physical forms and of human and machine-readable languages. Thomas Baker and Praditta Siripan describe how a diverse community of scientists, librarians, and bureaucrats have achieved agreement on the Dublin Core -- a general method for describing these objects in ways that everyone can use and understand. Putting such efforts into a broader historical context, Shigeo Sugimoto and Jun Adachi explain how librarians and information providers have built on and combined these technologies to develop projects in the emerging field of digital libraries.

The papers in the scientific workshop report on more specific topics and developments. Edie Rasmussen draws lessons from her experience in building a digital library to support elementary school teachers. Fytton Rowland considers the past few years of experimentation and offers some realistic reflections on how radically the new technologies may change -- or not change -- the nature of research libraries. Yoshikatsu Nagata describes the technology used to retrieve images in a multimedia digital library.

Following up on the tutorials about metadata and information structure, several papers address technical and strategic aspects of the movement towards new international standards in these areas. Eric Miller and Stuart Weibel provide an update on the rapidly evolving Dublin Core data model, while Thomas Baker relates this model to the activities of the Working Group for Dublin Core in Multiple Languages. Han Suk Choi describes how Dublin Core can be used in union catalogs of scholarly materials. Shigeo Sugimoto reports on a method for typing and searching text in languages like Japanese using Java applets on "off-the-shelf" Web browsers. Vilas Wuwongse contributes a critical assessment of the Resource Description Framework as a whole.

One major area of digital library research seeks to provide computers with the ability to recognize and display different writing systems and the intelligence to translate texts and search requests between languages. Akiko Aizawa describes new methods for generating subject keywords in multiple languages automatically. Asanee Kawtrakul discusses the specific linguistic problems encountered in searching through digital texts in the Thai language.

The papers in these proceedings reflect work in progress on a number of fronts. Presentations that reported on late-breaking news are represented here by their abstracts.

Thomas Baker
Asanee Kawtrakul
Program Co-Chairs

IJWDL98 Organization

Organizing Committee

Co-Chairs:

Vilas Wuwongse.(AIT, Thailand)
Chutima Sacchanand (TLA, Thailand)

Program Committee

Co-Chairs:

Thomas Baker (AIT, Thailand)
Asanee Kawtrakul (KU, Thailand)

Committees:

Sugimoto Shigeo (ULIS, Japan)
Koichi Tabata(ULIS, Japan)
Surapan Meknavin (NECTEC, Thailand)
Tassana Hanpol(TLA, Thailand)
Somsuang Prutikul(TLA, Thailand)

AIT: Asian Institute of Technology (<http://www.ait.ac.th>)

TLA: The Thai Library Association Under the Royal Patronage of
H R H Princess Maha Chakri Sirindhorn

KU: Kasetsart University (<http://www.ku.ac.th>)

ULIS: University of Library and Information Science (<http://www.ulis.ac.jp>)

NECTEC: National Electronics and Computer Technology Center
(<http://www.nectec.or.th>)

Table of Contents

Tutorials

Computer Technology and Networks for Digital Libraries	1
Nobuhito Yamamoto (University of Tsukuba, Japan)	
SGML/XML Re-Inventing the WEB	6
Masahiro Kikuta (Synergy Incubate Inc., Japan)	
Specification of XML	23
Youichi Shibata (Synergy Incubate Inc., Japan)	
Overview and Management of Digital Libraries II	40
A Case of NACSIS-EIS and Its On-line Journals	
Jun Adachi (National Center for Science Information Systems, Japan)	

Workshop

A Digital Library for K-12 Educators: the PEN-DOR project	46
Myron Bright, Karen Fullerton, Jane Greenberg, Maureen McClure, Edie Rasmussen and Darin Stewart (University of Pittsburgh, USA)	
Recent developments in scholarly publishing and their impact on libraries	47
Fytton Rowland (Loughborough University, UK)	
Retrieval System for the Microfilm Image Databases in the Media Center,	64
Osaka City University	
NAGATA Yoshikatsu, SHIBAYAMA Mamoru, KITA Katsuichi, MAEDA Harumi (Osaka City University, Japan)	
The Dublin Core Data Model	71
Eric Miller and Stuart Weibel (OCLC Office of Research, USA)	
Cataloging for the Web: Dublin Core and the Resource Description Framework ...	72
Thomas Baker (Asian Institute of Technology, Thailand)	
Introduction to DC-based Union Cataloging Systems for Academic	73
Journals in Digital Library Environments	
Han Suk Choi (Korea Research Information Center, Korea)	

Extension of MHTML to Text Input and Text Search Functions in Multiple Languages on Off-the-shelf Browsers	74
Shigeo Sugimoto ¹ , Shigetaka Nakao ¹ , Myriam Dartois ¹ , Jun Ohta ¹ , Akira Maeda ² , Tetsuo Sakaguchi ¹ , Koichi Tabata ¹ (University of Library and Information Science ¹ , Nara Institute of Science and Technology ² , Japan)	
Reasoning about RDF Elements	82
Vilas Wuwongse ¹ , Chutiporn Anutariya ¹ and Ekawit Nantajeewarawat ² (Asian Institute of Technology ¹ , Thammasat University ² , Thailand)	
A Graph-based Method for Automatic Generation of Multilingual Keyword Clusters and Its Applications	94
Akiko Aizawa, Noriko Kando and Kyo Kageura (National Center for Science Information Systems, Japan)	
An Intelligent Search for Thai Text in Digital Libraries	104
Asanee Kawtrakul, Thanussak Thanayasiri, Chavalit Chiraratachan, Nathavit Buranapraphanont, Preeti Piti-alongkorn, Navapt Khantonthong, Soontharee Koompairojn (Kasetsart University, Thailand)	
Author Index	105
Keyword Index	106

TUTORIALS

IJWDL'98

International Joint Workshop on Digital Libraries 1998

7-9 September 1998

Asian Institute of Technology

Bangkok, Thailand

Computer Technology and Networks for Digital Libraries

Nobuhito Yamamoto

Institute of Information Science and electronics

University of Tsukuba

Tennoudai 1-1-1, Tsukuba, Ibaraki 305-8573, Japan

1. Architecture of Computers

A computer system generally consists of four major functional blocks, processing unit, storage unit, input/output channel and peripheral devices. These blocks are connected with each other by a bus.

(1) Processing Unit (PU)

Numerical calculation operations such as addition, multiplication and comparisons are the basic operation inside a processing unit. Control of a program execution path like branching is also the important operation in it. After the invention of microprocessor technology, PU has been able to be manufactured in relatively lower cost. This caused the personal computer very popular nowadays.

(2) Storage Unit (SU)

The main storage or main memory is the information storage which PU can handle directory in fast speed. The memory is accessed randomly by its memory address. DRAM technology gives us the big mass capacity recently, e.g. 128Mbytes or 256Mbytes.

(3) Input/Output Channel

It is the interface between a main function box of computer and its peripheral devices. Information is read-in and written-out through I/O channel.

(4) Peripheral Devices

A computer works with numbers of peripheral devices, keyboard, display, pointing device such as a mouse and secondary storage units. Hard disks are the main devices of today's secondary storage unit. Floppy diskette drive, tape drive, magneto optical disk (MO) drive and compact disk (CD) drive are used for the secondary storage unit. Keyboard and display are the basic devices when a user uses a computer. Network connection device is increasing its role.

2. Hardware and Software

Not only hardware devices or boxes are a computer system. Software is the important partner components for hardware. Software is a series of processing instructions for

computers. We can not do any computation or processing without any software.

3. Operating System (OS)

An operating system is a special program or software which maintain the function of a computer and manages every computational processes including I/O. User's program or application works with the assistance of operating system. MacOS and Windows are the popular OS's for personal computers. UNIX is a very common OS on workstations. Conventional main frame computers have their own OS's. OS/MVS for IBM machines and GCOS for GE machines are well known.

4. Recent Trends of Computation

User-friendliness and high performance computing are the keywords for recent computation scene. A computer was before a physically big machinery which was installed at particular institutions like computer centers. That means users had to share the centralized computing resources. After the small scale computers advent, we have been able to have computing resources personally at our own places. And computers became a useful tool not only for the professional but for common people named end-users. User-friendliness is so one of the "must" condition for non professional end-users to use a computer. Graphical Users Interface (GUI) technology has brought easy understanding of situation a user is facing in front of a computer, say What you see is what you get (WYSIWYG).

High performance computing is the resent demand too. Traditional super computer has been using the vector processing technology which computes a series of the same kind of calculation operation in very high speed. Parallelism is another paradigm for increasing speed. And we can compute the same type of calculation concurrently using a number of processing units. Computers which have even thousands of PU have been manufactured now.

5. Networking

Networking of computers began at the end of 1950s in the United States. A handling unit of information called "packet" was introduced there, and packet transfer became the basic scheme of communication on the computer network. The first network was a mere set of connections among host computers installed at distant places. In 70s the concept of Local Area Network (LAN) was introduced at Xerox PARC. The UNIX development team at UC Berkeley chose ARPA network protocol known as TCP/IP for handling their LAN. The concept of networking was expanded to connections among

networks. This network has evolved to the Internet of today.

6. OSI Reference Model

The OSI reference model has been used frequently for explaining network structures. The model consists of seven logical layers. Though actual implementation of networking software may differ from the OSI model.

Communication between two hosts is performed with operations of going up and down the layers.

7. Transmission Media

Conventional serial lines and Ethernet (10Mbps) are the popular media for networking at the first generation. As the strong demand for high speed transmission, the Fast Ethernet and FDDI have been becoming the succeeding main media for LAN, whose transmission speed is 100Mbps. Asynchronous Transfer Mode (ATM) technology is also used for the latest network. Digital Subscriber Line (xDSL) technology is also the hot topic in this field.

8. IP address and MAC address

Every host on the TCP/IP network has its own individual IP address. Based on this address, a target host is identified whether it is on the same local network or is located on another network. Two hosts on the same network communicate using Media Access Control (MAC) addresses, which is associated to the IP address. This association is performed using one protocol in the TCP/IP suit.

9. Routing

At the gate of a network, information packets are handled whether they may pass the gate, and determined which interface should be chosen to the destination network. It is called "routing". Special hardware which performs routing is named a router. So a router connects adjacent (sub)networks and transfers packets from one network to another. Information for routing packets is exchanged and shared with each routers using routing protocols. RIP, OSPF and BGP are dominant and frequently used protocols.

10. Security

Securities are very important for both computers and networks. Personal computers are basically used as PERSONAL tool. That means one owner only use it and no other

person touches it. But for the common information resources such as database servers and computation servers, being secure is one of the most important requirements. The same thing can be said about the network. If a router is intruded, traffic of the network will be affected seriously and communication between hosts on that network will be hard to be established.

11. Applications

A computer can be applied solo in many jobs. Numerical computation or analysis was the most fundamental usage of computers. But recently the application field has been expanding and the usability become the main concern of users. Word processing and spread sheet processing are the most popular applications. After networking computers, usability has been increased incredibly. Below is some examples of typical network computing application.

11.1. Remote Resource Access

A computer connected to a network can be accessed from practically any places. Users can compute their own job using a host at a distance, or can share information stored on one computer. Information retrieval or database manipulation is the typical application of this kind at the libraries.

Handling old valuable documents in forms of digital photo images may be a convenient method which microfiches had been used for. Moreover, when all information on papers are converted to the electrical form, people will seemingly no more necessary to go to the place where that document is stored actually. Instead, he/she will use a network connection and access the information at their local places.

11.2. Multimedia Communication

As the development of network technology, we are now able to use computers/networks as useful media, in particular so called multimedia which transfer audio visual information.

World Wide Web (WWW) is one of the most successful and fruitful application. It handles hyper text documents very easily and give the facility of publication. WWW provide a good human interface for understanding the situation in front of him/her (Push technology).

Video Conferencing is also the most advanced and expectative application of network computing. We can communicate with the participants employing video images and voices.

Broadcasting on the network is a hot topic of multimedia application. Multicast transmission is a new technology which many network researchers and engineers are developing. It provides a bandwidth efficient way of transmission and therefore suits for the broadcasting purpose.

12. Conclusion

It is not necessary to say that the traditional library work is very important. For, that work handles the primary data directory. On that fundamental work, new technology will be expected to give us new aspects. Database, retrieval engine, GUI and networking, these will surely become much and more important keywords of the tomorrow's library activities.

References

- (1) Asynchronous Transfer Mode (ATM), Cisco technology directions, Cisco Systems, 1993.
- (2) Eriksson, H., Mbone: The Multicast Backbone, CACM, Vol.37, No.8, pp61-66, 1994.
- (3) Lynch, D.C., Rose, M.T., Internet system handbook, Addison-Wesley, 1993.
- (4) Page, N., White Paper, Streaming Web Video, Multimedia Access Corporation, 1997.
- (5) Thaler, D., et. al., Interoperability Rules for Multicast Routing Protocols, Internet-Draft, IETF, 1996.
- (6) VIA 188 - Distance Learning and Video Conferencing, AG Communication Systems, 1997.
- (7) Yamamoto, N., Multicast Communication on a Campus Environment, Proceedings of the 3rd International Workshop on Academic Information Networks and Systems, 1997.
- (8) Yamamoto, N., Communication among Sparse Groups using Streaming Video Technology, *Proceedings of the 4th International Workshop on Academic Information Networks and Systems*, 1998.

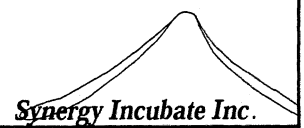


SGML/XML

Re-Inventing the WEB

8 Sept. 1998

Masahiro Kikuta
Synergy Incubate Inc.

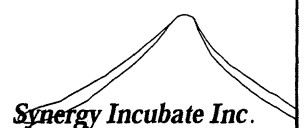


What is happening on the NET?

✓ **The extraordinary growth of the World-Wide Web**

<Key Technology>

- HTTP: Hypertext Transfer Protocol
- URL: Uniform Resource Locator
- HTML: Hypertext Markup Language



HTTP

✓ HTTP. Hypertext Transfer Protocol

- The client-server TCP/IP protocol used on the World-Wide Web for the exchange of HTML documents.

Synergy Incubate Inc.

URL

✓ URL:Uniform Resource Locator

- A draft standard for specifying an object on the Internet, such as a file or a newsgroup. URLs are used extensively on the World-Wide Web. They are used in HTML documents to specify the target of a hyperlink.

Synergy Incubate Inc.

.....
DR

HTML

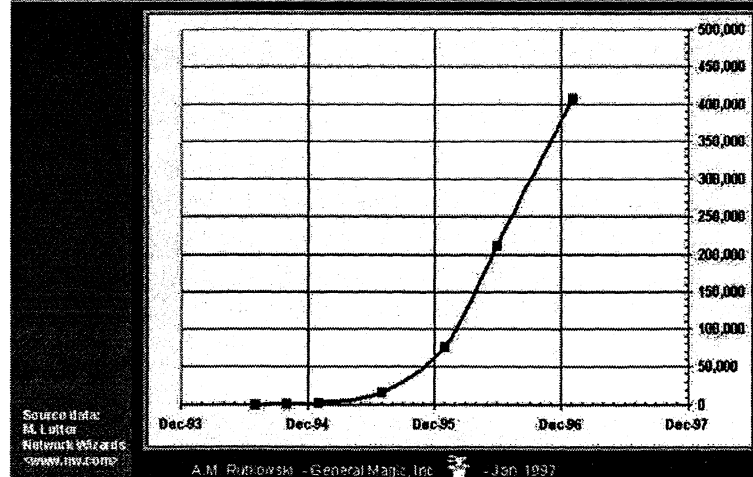
✓ HTML: Hypertext Markup Language

- A Hypertext document format used on the World-Wide Web. Built on top of SGML. "Tags" are embedded in the text. A tag consists of a "<", a "directive" (case insensitive), zero or more parameters and a ">". Matched pairs of directives, like "<TITLE>" and "</TITLE>" are used to delimit text which is to appear in a special place or style.

Synergy Incubate Inc.

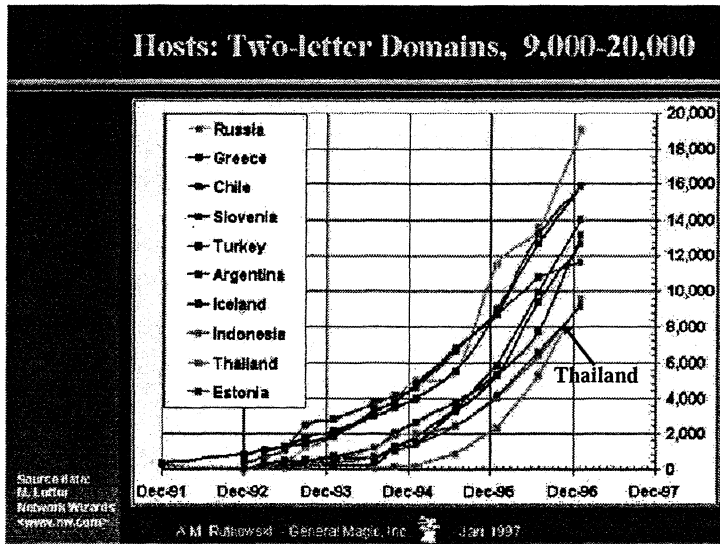
.....
DR

WWW-Prefixed Hosts



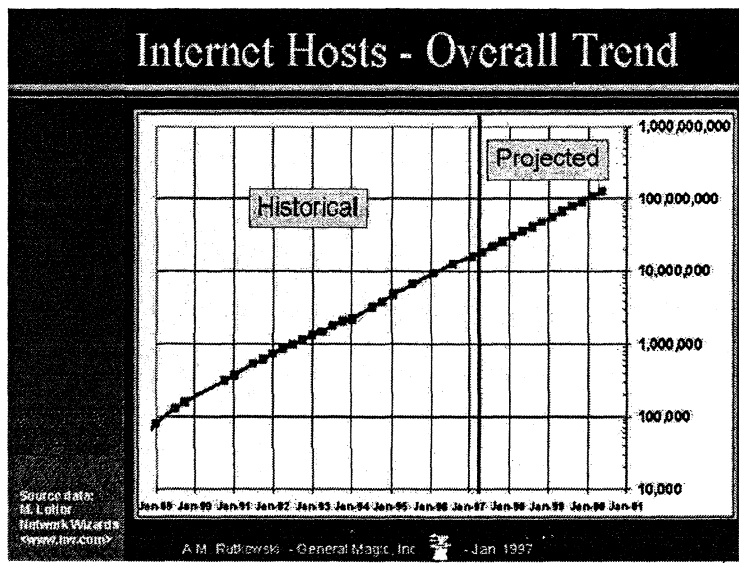
Synergy Incubate Inc.

DR



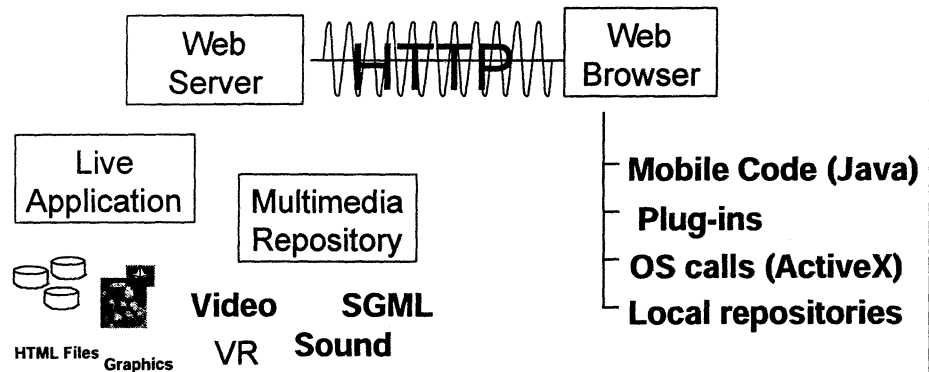
Synergy Incubate Inc.

DR



Synergy Incubate Inc.

The World-Wide Web: Multidimensional Multimedia



Synergy Incubate Inc.

What's wrong with HTML?

✓ HTML was optimized for easy learning

- One tag set for all applications
- Predefined semantics for each tag
- Predefined data structures
- No formal validation

✓ HTML trades power for ease of use

Synergy Incubate Inc.

..... 

What's wrong with HTML?

- ✓ **HTML is well suited to simple applications,**
- ✓ **But poorly suited to more demanding applications**
 - Large or complex collections of data
 - Data that must be used in many ways
 - Data with a long life cycle


Synergy Incubate Inc.

..... 

SGML offers,

- ✓ **Extensibility**
 - Authors can define new tag names and attribute names for documents by specifying their their syntax and semantics.
- ✓ **Structure**
 - Documents can be containers for other documents, with arbitrary nesting.
 - This allows complex documents to be constructed from simpler documents.


Synergy Incubate Inc.

.....
JK

SGML offers,

✓ Validation

- If desired, any SGML document can reference a description of its grammar so applications can validate that the document conforms to its specified structure.

IEEE Internet Computing
By Rohit Khare/ Adam Rifkin


Synergy Incubate Inc.

.....
JK

SGML can be

- ✓ Safest of all formats for high-value data (ISO standard 8879)
- ✓ Best available format for search/retrieval.
- ✓ Easily converted to HTML.

But.....


Synergy Incubate Inc.

What' wrong with SGML

- ✓ Designed before the PC revolution.
- ✓ Large, cumbersome specification (500pages)
- ✓ Difficult for people to read and understand
- ✓ Difficult for computers to process and manipulate
- ✓ High entry cost

Synergy Incubate Inc.

Motivation to XML

- ✓ SGML has been here since 1986, and
- ✓ Many of us think it's a good idea, but
- ✓ It hasn't caught on that fast, and
- ✓ The Web is here, and
- ✓ We need it even more than ever!

Synergy Incubate Inc.



What is XML?

- ✓ **Next step in Web evolution**
- ✓ **Goes beyond the limitations of HTML**
- ✓ **Will create new Web applications**
 - database exchange
 - Distribution of processing to clients
 - Client-side manipulation of views into the data
 - Customization of information by intelligent agents
 - Management of document collections

Synergy Incubate Inc.



Why XML?

- ✓ **For a new generation of Web applications:**
 - ***Extensibility:*** Users can define new tags as needed
 - ***Structure:*** Hierarchical data can be modeled to any level of complexity
 - ***Validation:*** Data can be checked for structural correctness
 - ***Media independence:*** The same content can be published in multiple media


Synergy Incubate Inc.

.....



XML Design Goals

- 1.XML shall be straightforwardly usable over the Internet**
- 2.XML shall support a wide variety of applications**
- 3.XML shall be compatible with SGML**
 - Existing SGML tools read & write XML
 - XML documents are SGML documents
 - The same parse can be generated
 - Similar expressive power




Synergy Incubate Inc.

.....



XML Design Goals

- 4.It shall be easy to write programs which process XML documents.**
- 5.The number of optional features in XML is to be kept to the absolute minimum, ideally zero**
- 6.XML documents should be human-legible and reasonably clear**



Synergy Incubate Inc.

.....



XML Design Goals

- 7. The XML design should be prepared quickly**
- 8. The design of XML shall be formal and concise**
- 9. XML documents shall be easy to create**
- 10. Terseness is of minimal importance**



Synergy Incubate Inc.

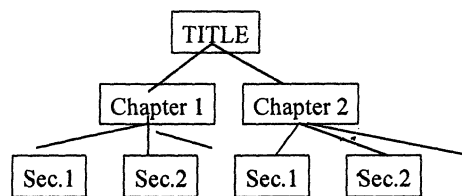
.....



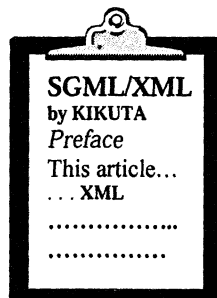
Document is composed of

These slides are designed for XML tutorial seminar in IJWDL '98,.....

CONTENT



STRUCTURE



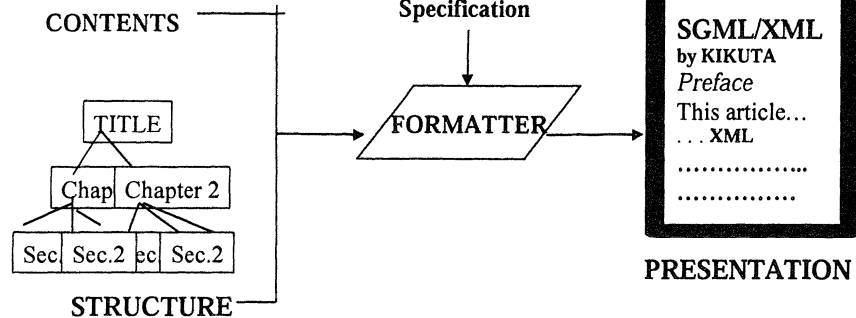
PRESENTATION



Synergy Incubate Inc.

XML allow us:

These slides are designed for XML tutorial seminar in IJWDL'98,.....



XML specifies the content and structure of document in a way that allows particular presentation.

Synergy Incubate Inc.

Relationship to SGML

- ✓ XML is a simplified subset of SGML:
 - Powerful
 - No limits on namespace of structural depth
 - Easy to implement
 - Small enough for Web browsers
- ✓ The translation from SGML to XML is straightforward

Synergy Incubate Inc.

..... DR

What is the role of XML?

✓ SGML


- For large publication systems
- High-value information

✓ XML

- For structured delivery over the web
- In many cases will be adequate for authoring

✓ HTML

- For delivery over the web
- Low-value information

 Synergy Incubate Inc.

..... DR

Advantages of XML

- ✓ **Conformant with existing international standard (SGML)**
- ✓ **Complete extensibility -- no tag limitations**
- ✓ **Full internationalization**
- ✓ **Validation and editorial control**
- ✓ **Ability to model any kind of hierarchical data**

 Synergy Incubate Inc.



Advantages of XML

- ✓ **Automatic generation of links and navigational aids**
- ✓ **Increased speed of access to essential data**
- ✓ **Print and online versions from same source**
- ✓ **Dynamically user-configurable views**
- ✓ **Next-generation hypertext capabilities**

Synergy Incubate Inc.



The XML family

- ✓ **XML (eXtensible Markup Language)**
 - A subset of SGML(ISO 8879) designed for easy implementation
- ✓ **XLL(eXtensible Linking Language):**
 - A set of standard hypertext mechanisms based on HyTime(ISO/IEC 10744)and the Text Encoding Initiative(TEI)

Synergy Incubate Inc.


.....



The XML family

✓ XSL(eXtensible Stylesheet Language)

- A standard stylesheet language for structured information formed by subsetting DSSSL (ISO/IEC 10179), designing an alternative syntax, and incorporating key CSS concepts



Synergy Incubate Inc.

.....



Industry specific tag language

- ✓ **CML:Chemical Markup Language**
- ✓ **CDF:Channel Definition Format**
- ✓ **OFX:Open Financial Exchange**
- ✓ **HDML:Handheld Device Markup Language**
- ✓ **MathML:Mathematics Markup Language**
- ✓ **PGML:Precision Graphics Markup Language**



Synergy Incubate Inc.



The XML applications can

- ✓ use Web clients to mediate between multiple heterogeneous databases
- ✓ distribute the load from Web servers to their clients
- ✓ use Web clients to present different views of the same data
- ✓ employ agents to tailor information discovery and filtering to the customized needs

Synergy Incubate Inc.



User view of the data


- ✓ **Different views of whole documents**
 - Novice view vs. expert view
 - Outline vs. content
 - Generated Tables of Contents
- ✓ **Different views of document components**
 - Bar graph vs. pie chart
 - Totals by region vs. totals by business units

Synergy Incubate Inc.

..... 

XML require additional mechanisms

- ✓ Programs, applets, or scripts designed for a specific tag set
- ✓ Industry agreements on the processing of specific tags
- ✓ Tag-sensitive components.
- ✓ Stylesheets



Synergy Incubate Inc.

..... 

The future of the Web


- ✓ View Selection: Letting The User Decide
- ✓ Web Agents: Data That knows About Me
- ✓ Users no longer tied to a proprietary format
- ✓ An end to domination of the market by a few big companies
- ✓ An end to domination of the market by a few big countries

Jon Bosak Chairman of XML WG, W3C


Synergy Incubate Inc.

› Specification ‹ of XML

- Extensible Markup Language (XML) 1.0 -

Youichi Shibata Synergy Incubate Inc. 

› XML documents ‹

- Legal XML documents are said to be well-formed having a each logical tree with a single root
- The XML documents with an optional set of constraints (a DTD), they are said to be valid

Synergy Incubate Inc. 

> **Well-formed Tag/Attribute Syntax** <

- . Element types and attribute names are case-sensitive
- . Empty elements must be marked by closing />
- . Attribute values must be quoted (' or ")

Synergy Incubate Inc. 

> **Element of XML Documents** <

- . XML Documents are composed of;
 1. prolog
 2. element
 3. Misc

[1] document ::= prolog element Misc*

[27] Misc ::= Comment | PI | S

Synergy Incubate Inc. 

› XML Declaration ‹

1. Specify of Version Number
2. Encording Declaration
3. Standalone Document Declaration

example:<?xml version="1.0" encoding="UTF-8"
standalone="no" ?>


[23] XMLDecl ::= '<?xml' VersionInfo EncodingDecl?
SDDDecl? S? '?>'

Synergy Incubate Inc. 

› Encording Declaration ‹

1. UTF-8
2. UTF-16
3. ISO-10646-UCS-2
4. ISO-10646-UCS-4
5. ISO-8859-1
6. ISO-8859-2
7. :
8. :
9. ISO-8859-9
10. Shift-JIS
11. EUC-JP
12. ISO-2022-JP

[80] EncodingDecl ::= S 'encoding' Eq (' ' EncName ' ' | ' ' ' EncName ' ' ')

Synergy Incubate Inc. 

› Standalone Document Declaration ◀

- The value "yes" indicates that there are nomarkup declarations external to the document entity
- The value "no" indicates that there are or may be such external markup declarations

* Load reduction for XML processors

```
[32] SDDecl ::= S 'standalone' Eq (('"' ('yes' | 'no')  
    """) | ('"' ('yes' | 'no') '"'))
```

Synergy Incubate Inc. 

› Document Type Declaration ◀

- The XML document type declaration contains or points to markup declarations

```
[28] doctypedekl ::= '<!DOCTYPE' S Name  
    (S ExternalID)? S? ('[' (markupdecl|PEReference|S)*  
    ']' S?)? '>'
```

```
[75] ExternalID ::= 'SYSTEM' S SystemLiteral  
    | 'PUBLIC' S PubidLiteral S SystemLiteral
```

```
[11] SystemLiteral ::= ('"' [^"]* '"')  
    | ('"' [^']* '"')
```

```
[12] PubidLiteral ::= '"' PubidChar* '"'  
    | '"' (PubidChar - '"')* '"'
```


```
[13] PubidChar ::= #x20 | #xD | #xA | [a-zA-Z0-9]  
    | [-'()+,./:=?;!*#@$_%]
```

Synergy Incubate Inc. 

› Markup Declaration ‹

- . Markup declaration provide a grammar of documents
- . DTD(document type definition) is a grammar of documents
- . Markup declaration declares;
 1. element type declaration
 2. attribute-list declaration
 3. entity declaration
 4. notation declaration
 5. Processing Instruction
 6. Comments


```
[29] markupdecl ::= elementdecl | AttlistDecl  
| EntityDecl | NotationDecl | PI | Comment
```

Synergy Incubate Inc. 

› Element ‹

- . Tow types Elements
 1. Empty-Element Tags
 2. Start-Tags,content,End-Tags

```
[39] element ::= EmptyElemTag | STag content ETag
```

Synergy Incubate Inc. 

› Empty-Element Tags ‹

- Empty-element tags may be used for any element which has no content, whether or not it is declared using the keyword EMPTY. For interoperability, the empty-element tag must be used, and can only be used, for elements which are declared EMPTY

example:

```
<IMG src="http://www.w3.org/Icons/WWW/w3c_home" />  
[44] EmptyElemTag ::= '<' Name (S Attribute)* S? '/>'
```

Synergy Incubate Inc. 

› Start-Tags, Content, End-Tags ‹

- The beginning of every non-empty XML element is marked by a start-tag.
- The Name in the start- and end-tags gives the element's type. The Name-AttValue pairs are referred to as the attribute specifications of the element, with the Name in each pair referred to as the attribute name and the content of the AttValue (the text between the ' or " delimiters) as the attribute value

example:


```
<aff>Synergy Incubate Inc.</aff>  
[40] STag ::= '<' Name (S Attribute)* S? '>'  
[43] content ::= (element | CharData | Reference  
| CDSect | PI | Comment)*  
[42] ETag ::= '
```

Synergy Incubate Inc. 

> Character data <

- All text that is not markup constitutes the character data of the document
- The ampersand character (&) and the left angle bracket (<) may appear in their literal form only when used as markup delimiters, or within a comment, a processing instruction, or a CDATA section
- The right angle bracket (>) may be represented using the string ">", and must, for compatibility, be escaped using ">" or a character reference when it appears in the string "]]>" in content, when that string is not marking the end of a CDATA section

[14] CharData ::= [^&]* - ([^&]* ']]>' [^&]*)

Synergy Incubate Inc. 

> Reference <

1. Entity Reference
 - An XML document may consist of one or many storage units. These are called entities
 - An entity reference refers to the content of a named entity. References to parsed general entities use ampersand (&) and semicolon (;) as delimiters
2. Character Reference
 - A character reference refers to a specific character in the ISO/IEC 10646 character set, for example one not directly accessible from available input devices

[67] Reference ::= EntityRef | CharRef

[68] EntityRef ::= '&' Name ';'

[66] CharRef ::= ' ' [0-9]+ ';' | '&#x' [0-9a-fA-F]+ ';' |

Synergy Incubate Inc. 


> CDATA Sections <

- CDATA sections may occur anywhere character data may occur; they are used to escape blocks of text containing characters which would otherwise be recognized as markup. CDATA sections begin with the string "<![CDATA[" and end with the string "]]>"

example:

```
<![CDATA[<greeting>Hello, world!</greeting>]]>
```

```
[18] CDsect ::= CDstart CData CEnd
[19] CDstart ::= '<![CDATA['
[20] CData ::= (Char* - (Char* ']]>' Char*))
[21] CEnd ::= ']]>'
```

Synergy Incubate Inc. 

> (PI) Processing Instructions <

- Processing instructions (PIs) allow documents to contain instructions for applications

example:

```
<?metaviewer color green ?>
```

```
[16] PI ::= '<?' PITarget
      (S (Char* - (Char* '?>' Char*)))? '?>'
[17] PITarget ::= Name - (('X'|'x') ('M'|'m') ('L'|'l'))
```

Synergy Incubate Inc. 

» Comments ◀

- Comments may appear anywhere in a document outside other markup; in addition, they may appear within the document type declaration at places allowed by the grammar. They are not part of the document's character data

example:

```
<!-- declarations for & -->
```

```
[15] Comment ::= '<!--' ((Char - '-')  
| ('-' (Char - '-')))* '-->'
```

Synergy Incubate Inc. 


» Document Type Definition (DTD) ◀

- XML documents may, and should, begin with an XML declaration which specifies the version of XML being used
- The document type declaration can point to an external subset (a special kind of external entity) containing markup declarations
- A markup declaration is an element type declaration, an attribute-list declaration, an entity declaration, or a notation declaration

```
[28] doctypedecl ::= '<!DOCTYPE' S Name  
(S ExternalID)? S? ('[' (markupdecl  
| PEReference | S)* ']' S?)? '>'
```

```
[29] markupdecl ::= elementdecl | AttlistDecl  
| EntityDecl | NotationDecl | PI | Comment
```

```
[69] PEReference ::= '%' Name ';' ;
```

Synergy Incubate Inc. 

› Element Type Declarations ‹

- . The element structure of an XML document may, for validation purposes, be constrained using element type and attribute-list declarations. An element type declaration constrains the element's content
- . Element type declarations often constrain which element types can appear as children of the element


Synergy Incubate Inc. 

› Element Content ‹

- . An element type has element content when elements of that type must contain only child elements (no character data)
- . An element is valid if there is a declaration matching elementdecl where the Name matches the element type, and one of the following holds
 1. The declaration matches EMPTY and the element has no content
 2. The declaration matches children and the sequence of child elements belongs to the language generated by the regular expression in the content model, with optional white space (characters matching the nonterminal S) between each pair of child elements
 3. The declaration matches Mixed and the content consists of character data and child elements whose types match names in the content model
 4. The declaration matches ANY, and the types of any child elements have been declared

```
[45] elementdecl ::= '<!ELEMENT' S Name  
S contentspec S? '>'
```

```
[46] contentspec ::= 'EMPTY'|'ANY'|Mixed|children
```

Synergy Incubate Inc. 

> Attribute-List Declarations <

- . Attribute-list declarations may be used
 - . To define the set of attributes pertaining to a given element type
 - . To establish type constraints for these attributes
 - . To provide default values for attributes

```
[52] AttlistDecl ::= '<!ATTLIST' S Name  
AttDef* S? '>'
```

```
[53] AttDef ::= S Name S AttType S DefaultDecl
```

Synergy Incubate Inc. 

> Attribute Types <

- . XML attribute types are of three kinds: a string type, a set of tokenized types, and enumerated types

```
[54] AttType ::= StringType | TokenizedType  
| EnumeratedType
```


```
[55] StringType ::= 'CDATA'
```

```
[56] TokenizedType ::= 'ID'|'IDREF'|'IDREFS'|'ENTITY'  
|'ENTITIES'|'NMTOKEN'|'NMTOKENS'
```

```
[57] EnumeratedType ::= NotationType | Enumeration
```

```
[58] NotationType ::= 'NOTATION' S  
'(' S? Name (S? '|' S? Name)* S? ')'
```

```
[59] Enumeration ::= '(' S? Nmtoken  
(S? '|' S? Nmtoken)* S? ')'
```

Synergy Incubate Inc. 

› Attribute Defaults ‹

- An attribute declaration provides information on whether the attribute's presence is required, and if not, how an XML processor should react if a declared attribute is absent in a document

```
[60] DefaultDecl ::= '#REQUIRED' | '#IMPLIED'  
| (( '#FIXED' S)? AttValue)
```

Synergy Incubate Inc. 

› Physical Structures ‹

- An XML document may consist of one or many storage units. These are called entities; they all have content and are all identified by name. Each XML document has one entity called the document entity
 1. General/Parameter Entities
 2. Internal/External Entities
 3. Parsed/Unparsed Entities

Synergy Incubate Inc. 

› Entity Declarations ‹

```
[70] EntityDecl ::= GEDecl | PEDecl
                        /* GEDecl:General Entities */
                        /* PEDecl:Parameter Entities */
[71] GEDecl ::= '<!ENTITY' S Name S
EntityDef S? '>'
[72] PEDecl ::= '<!ENTITY' S '%' S
Name S PEDef S? '>'
[73] EntityDef ::= EntityValue
| (ExternalID NDataDecl?)
[74] PEDef ::= EntityValue | ExternalID
```

Synergy Incubate Inc. 

› Internal/External Entities ‹

Example of an internal entity declaration
<!ENTITY Pub-Status "This is the specification.">

If the entity is not internal, it is an external entity

```
[75] ExternalID ::= 'SYSTEM' S SystemLiteral
| 'PUBLIC' S PubidLiteral S SystemLiteral
[76] NDataDecl ::= S 'NDATA' S Name
```


Examples of external entity declarations:

```
<!ENTITY open-hatch
SYSTEM
"http://www.synergy.co.jp/boilerplate/OpenHatch.xml">
<!ENTITY open-hatch
PUBLIC
"-//synergy//TEXT Standard open-hatch boilerplate//EN"
"http://www.synergy.co.jp/boilerplate/OpenHatch.xml">
<!ENTITY hatch-pic
SYSTEM "../grafix/OpenHatch.gif"
NDATA gif >
```

Synergy Incubate Inc. 

› **Parsed/Unparsed Entities** ‹

- Parsed Entities
Text Data, parsed as the ingredients of XML Documents
- Unparsed Entities
Entities that may not be parsed (Graphic and/or Text Data)

Synergy Incubate Inc. 

› **Entities may be used without declaration** ‹

- < : <
- > : >
- & : &
- ' : '
- " : "

Synergy Incubate Inc. 

› Entity Reference ‹

- . An entity reference refers to the content of a named entity. References to parsed general entities use ampersand (&) and semicolon (;) as delimiters
- . Parameter-entity references use percent-sign (%) and semicolon (;) as delimiters

[67] Reference ::= EntityRef | CharRef

[68] EntityRef ::= '&' Name ';' {&examples}

[69] PReference ::= '%' Name ';' {%examples}


Synergy Incubate Inc. 

› Notation Declarations ‹

- . Notations identify by name the format of unparsed entities, the format of elements which bear a notation attribute, or the application to which a processing instruction is addressed
- . Notation declarations provide a name for the notation, for use in entity and attribute-list declarations and in attribute specifications, and an external identifier for the notation which may allow an XML processor or its client application to locate a helper application capable of processing data in the given notation

[82] NotationDecl ::= '<!NOTATION' S Name S
(ExternalID | PublicID) S? '>'

[83] PublicID ::= 'PUBLIC' S PubidLiteral

Synergy Incubate Inc. 



White Space



- S (white space) consists of one or more space (#x20) characters, carriage returns, line feeds, or tab

```
[3] S ::= (#x20 | #x9 | #xD | #xA)+
```

Synergy Incubate Inc. 



Language Identification



- In document processing, it is often useful to identify the natural or formal language in which the content is written. A special attribute named xml:lang may be inserted in documents to specify the language used in the contents and attribute values of any element in an XML document

```
[33] LanguageID ::= Langcode ('-' Subcode)*
```

```
[34] Langcode ::= ISO639Code | IanaCode | UserCode
```

```
[35] ISO639Code ::= ([a-z] | [A-Z]) ([a-z] | [A-Z])
```

```
[36] IanaCode ::= ('i' | 'I') '-' ([a-z] | [A-Z])+
```

```
[37] UserCode ::= ('x' | 'X') '-' ([a-z] | [A-Z])+
```

```
[38] Subcode ::= ([a-z] | [A-Z])+
```

Synergy Incubate Inc. 

› Declaring Namespaces ‹

- XML namespaces are based on the use of qualified names, which contain a single colon, separating the name into a namespace prefix and the local name. The prefix, which is mapped to a URI, selects a namespace. The combination of the universally-managed URI namespace and the local schema namespace produces names that are guaranteed universally unique.

```
[1] NamespacePI ::= '<?xml:namespace'  
    (S (PrefixDef | NSDef | SrcDef))+ '?>'  
[2] NSDef ::= 'ns' Eq SystemLiteral  
[3] SrcDef ::= 'src' Eq SystemLiteral  
[4] PrefixDef ::= 'prefix' Eq  
    ('"' NCName '"' | "'" NCName "'")  
[5] NCName ::= (Letter | '_' ) (NCNameChar)*  
[6] NCNameChar ::= Letter | Digit | '.' | '-'  
    | '_' | CombiningChar | Extender
```

Synergy Incubate Inc. 

Overview and Management of Digital Libraries II

A Case of NACSIS-ELS and Its On-line Journals

Jun Adachi

to be presented at IJWDL'98, 8th September 1998, AIT, Thailand

Research and Development Department
NACSIS (National Center for Science Information Systems)

3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-8640, Japan

Abstract NACSIS initiated an Internet-based document delivery service called NACSIS-ELS (NACSIS Electronic Library Service) in April 1997. As of August 1998, 55 Japanese academic societies are participating in this project, and nearly 170 academic journals are captured and made available on NACSIS-ELS. The history of the development of NACSIS-ELS is described after explaining general activities of NACSIS, Japan. Then a copyright charging strategy in this new service is discussed. Other issues related to online journals are also mentioned such as security protection measures, academic society activities over the Internet.

1 Introduction

In this lecture, the author would like to introduce the activities of NACSIS in terms of online journal services and then to give an overview of current issues related to online journals, in particular, pricing strategies of various services.

2 Overview of NACSIS Activities

NACSIS, which was established in 1986, is an inter-university research institution under the auspices of the Japanese Ministry of Education, Science, Culture and Sports. The major role of NACSIS is to provide support for scientists and universities in terms of (1) information services, more specifically, compilation of scientific databases and services of online information retrieval from those databases, (2) services

related to university library networking and union catalog databases of books and serials, and (3) the management and operation of the Internet backbone for Japanese universities.

Since April 1997 the National Center for Science Information Systems (NACSIS) has been providing an electronic document delivery service called NACSIS-ELS (NACSIS Electronic Library Service). In this service Japanese academic journals are captured and made available to researchers through the Internet.

NACSIS-ELS, which was developed by the staff of NACSIS R&D Department, has a close relationship with the above mentioned NACSIS database activities.

2.1 Cataloging Services

One of the most heavily utilized services is NACSIS-CAT service, which is an online cataloging system for the compilation of the union catalogs of university library materials. The records of catalogs and holdings of books and serials at university libraries accumulate in databases day by day through the NACSIS-CAT service, which started in 1984. At present, more than 400 university libraries including 94 national universities, hundreds of private universities, some inter-university research institutions, and several public libraries are connected online to the NACSIS central catalog database server, which stores more than 20 million holdings records in total. These records are utilized for OPAC services at each university library as well.

2.2 Information Retrieval

The catalog databases made by libraries through NACSIS-CAT are offered in NACSIS-IR (NACSIS

Information Retrieval service), which is an online information retrieval service largely for scientists and has been operational since 1987. For example, users can access NACSIS-IR by telnet and enjoy more than 50 scientific and scholarly databases.

2.3 Document Delivery

The service which combines catalog information and general databases is NACSIS-ILL (NACSIS Inter-Library Loan service), which runs a kind of e-mail system for exchanging inquiries for document delivery from users to libraries. Users can issue a request to receive photocopies of an article which he or she has found by using NACSIS-IR. The request will be sent to one of the libraries that hold the journal or the proceedings that carries the article.

University libraries have been providing photocopying services of materials for years, and NACSIS-ILL has boosted the usage of this service by integration with online databases.

3 Cooperation with Academic Societies

3.1 Japanese Academic Societies

A major reason for government support of the activities of academic societies through NACSIS operations is that in Japan the scale of individual societies is small compared with those of the U.S., even though the number of societies is more than one thousand. Therefore, the financial state of the societies is weak, leading to occasional difficulties in publishing materials. Furthermore, Japanese societies are responsible for scholarly publications in the Japanese language, and those outside Japan do not subscribe easily to these publications. This makes societies financially weaker in the age of the Internet, where English prevails. While some societies publish journals in English, journals with an international reputation are few and most have limited subscriptions from overseas.

3.2 NACSIS and Societies

NACSIS is supporting the activities of Japanese academic societies by compiling bibliographic records of technical reports and conference papers that cannot be obtained easily through the usual publication distribution systems. NACSIS considers that gray literature,

(i.e., literature outside the usual distribution market, and difficult for people to obtain easily, such as SIG reports and conference proceedings) is of as much importance for researchers and scientists as academic periodicals.

Therefore, NACSIS initiated the bibliographic database compilation of Japanese gray literature. These databases, which include information such as paper title, author names, affiliation, and abstracts, are mounted in NACSIS-IR, the online information retrieval service provided by NACSIS. At present, 63 societies are participating in this database compilation.†

NACSIS has also initiated the compilation of full-text records of articles in some scientific journals in cooperation with the publishing departments of several societies.‡

NACSIS proposes to use SGML format for full-text encoding, but the number of participating journals is limited at present. In addition, it must be admitted that SGML database compilation is still at an experimental stage in Japan.

However, Japanese academic societies well understand that the trend is towards the electronic production of scientific journals. NACSIS is strongly promoting this in order to expand the compilation of primary information databases in science, technology and the humanities.

4 NACSIS-ELS Service and Its functions

NACSIS-ELS is one of new information services planned by NACSIS for academic communities.

NACSIS considered that the next stage for the further dissemination of scholarly information was an electronic journal service for Japanese academic societies. In this context, NACSIS-ELS was developed to accumulate scholarly information such as machine-readable journal articles, proceedings and technical reports.

The functions now available on NACSIS-ELS are online document delivery capabilities on the Internet. NACSIS-ELS has unique features as compared with other digital library projects which are based on conventional libraries. For example, the coverage of materials provided by NACSIS-ELS is restricted to scientific and scholarly journals, and potential users are assumed to be mainly scientists. It has no paper-form materials and permits access for users only through networks.

4.1 Databases

NACSIS-ELS accumulates mainly scientific and scholarly information, such as machine-readable jour-

† Negishi, M., "Databases of NACSIS (in Japanese)," *Joho-Shori*, Vol. 33, No. 10, pp. 1144-1153 (1992).

‡ Negishi, M., "Recent development in full text database applications (in Japanese)," *Joho-Shori*, Vol. 33, No. 4, pp. 413-420 (1992).

nal articles, proceedings and technical reports. Firstly, the database server stores conventional bibliographic databases that contain records comprising titles, author names, affiliations and abstracts of scientific papers. A catalog database of journals is also available. These databases are basically the same as those provided on NACSIS-IR.

As well as these conventional databases, the server holds page image databases of scientific journals, which include all pages from cover page to back cover. Pages are captured in a raster image digital format. 400 dpi resolution data is employed for page printouts, which is better quality than usual photocopies.

Journals are acquired from Japanese academic societies that have given NACSIS permission to use their journals for the NACSIS-ELS service.

4.2 Functions

Documents stored in the databases are retrieved and transmitted through high-speed wide-area networks and users can browse articles on their workstation monitors and users can print papers on a nearby high-quality laser printer, if necessary. Therefore, this system is a superset of conventional library services such as document delivery and photocopying services, and will soon supplant those conventional services.

The central server provides online access to end-user workstations through the Japanese Internet, such as SINET and university campus networks.

Transmission of images over a wide-area network requires wider bandwidth than text or audio information due to the large size of image data. NACSIS's SINET, which is one of the Internet backbones in Japan, was equipped in 1995 with ATM switches and experimental 155-Mbps fiber-optic circuits, which form the test-bed for a next-generation high-speed network.

4.3 Page Browsing

Two versions of client software are available for browsing pages of NACSIS-ELS. One is a Z39.50 based dedicated browser, and the other is a plug-in that works with conventional WWW browsers, such as Netscape Navigator. Both clients are distributed free of charge over the Internet.

With the Z39.50 browser, users can register journals they are interested in as their own environment setting in the user profile. First the server sends thumb-nail displays of journal cover registered in the user profile. Clicking on one of the cover page icons on the monitor retrieves the actual cover page of the latest issue of the specified journal, and allows free

browsing of pages. A hyperlink mechanism between contents entries and corresponding pages is embedded to allow access to the first page of the requested article from the contents.

4.4 Keyword Search

On the other hand, the user can initially search for papers in the same way as in conventional online information retrieval systems, i.e., using a keyword search. Then the server will send the short informations for articles that satisfy the query. After selecting an article to browse, its pages can be displayed in a window. Users can retrieve documents using bibliographic information such as titles, author names, or abstracts. Users can also issue more complex queries including boolean operators and result sets operations.

NACSIS-ELS is therefore unique in that it provides the usual functions of the online information retrieval systems found in NACSIS-IR, while also enabling users to browse pages on a monitor and then print those of interest.

A typical usual workstation monitor displays a whole A4 page in 75 dpi (dots per inch) resolution. This is not good enough for perusal of pages. Thus, an enlargement capability is embedded in the client software as well as page browsing functions to jump to next, previous, and cover pages.

4.5 Z39.50 Standards

To exchange database records via a network, NACSIS-ELS employs an extended version of ANSI Z39.50 (Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection), a standard by the National Information Standards Organization (NISO) which was designed as the standard protocol for information retrieval.

The original version of ANSI Z39.50 had to be extended to handle image data, thereby making compatibility and coordination with other Internet information services a crucial issue. The next version of NACSIS-ELS software, which will be released at the end of 1998, is expected to adopt Z39.50 Version 3 protocol, which will allow NACSIS-ELS users to access other Z39.50 servers on the global Internet as well. NACSIS-ELS also complies with ISO 8777 (Z39.58) in query commands.

4.6 Hyper-links between Page Images

A general mechanism of hyper-links between pages is to be embedded to allow users to jump from page to page by clicking mouse buttons. This mechanism allows the user to jump from a contents entry or a

bibliographic reference to the corresponding page, using the page link information embedded in rectangle areas on the pages. Since data for hyper-links are created manually at present, it is difficult to implement this function on all pages for the time being. The extraction of rectangle areas on a page image and application of OCR techniques would automate the generation of page links and we are now considering incorporating this capability into NACSIS-ELS software.

5 Brief History of Development

5.1 Development

NACSIS R&D Department started the development of NACSIS-ELS software in the early 90s. The first prototype was completed in 1993.

The Information Processing Society of Japan (IPSJ), a leading Japanese academic society in the field of computer and information science and other 2 societies graciously allowed NACSIS in 1993 to capture and store images of pages from their journals under restricted conditions. Work is proceeding on image database compilation of all IPSJ publications.

5.2 Trial Service

To evaluate the system performance of NACSIS-ELS, trial service was started in February 1995. Under it, selected users are provided with online access to 56 titles of journals through SINET. Based on user evaluations of system functionality and performance, the system is being developed further in an effort to offer a fully functional service. More than 300 users were participating in the evaluation of NACSIS-ELS. In the trial service, no copyright charges are applied to monitor users. This trial ended in March 1997.

5.3 Service Operation

NACSIS launched the full service of NACSIS-ELS in April 1997.

As of August 1998, about 7000 users are registered to the server. Fifty-five societies are participating in NACSIS-ELS, and other 56 societies are under consideration for participation. Nearly 170 journals will be digitized. Most of them will be soon available on the Internet this year. We have already digitized over 1 million pages. In 1998, additional 500 thousand pages are to be digitized. Since discussion with societies on copyright charging issues is still going on, there is no charge for the time being. NACSIS-ELS intends to institute a system of charges from January 1999.

Concurrently, retrospective digitization is also being experimented with. In 1995, all the pages of the Journals of Japanese Society for Artificial Intelligence were digitized from the first issue of the Journal in celebration of the Society's 10th anniversary. Similar projects are planned for other societies.

6 Copyright Issues in NACSIS-ELS

The definition of the copyright charging policy is one of the major concerns of academic societies. Efforts have been made to define a desirable policy. NACSIS is talking with not only academic societies but also with the organizations related to copyright issues and copyright clearance, expecting to establish a harmonized scheme on copyright charges for scholarly information in case of online dissemination. As of April 1998, negotiations have been settled on the whole.

6.1 Overview

Copyright processing, by which fees are collected from users and divided among the participating copyright holders, i.e., the academic societies, is closely linked to service policies and technical design issues. NACSIS has been negotiating with societies for some years and a general scheme is already established; current discussions will determine the details of the scheme and the price setting for each service item.

NACSIS-ELS permits users to display and printout journal pages, and the central server detects each user action such as page display and printout, and logs detailed usage information. The copyright charge will be calculated periodically based on the statistical processing of the log records. The items to be charged and their unit charges are now under negotiation.

In the next section, issues concerning price setting are described.

6.2 Concerns of Academic Societies

In summary, the concerns of societies in regard to NACSIS-ELS are the potentialities of (1) losing journal subscriptions and (2) a decrease in members because of far easier access to the information produced by the societies.

However, NACSIS anticipates that NACSIS-ELS will boost the activities of academic societies without negative influences and is continuing to discuss the price setting strategy with the societies involved.

6.3 Principles

The copyrights of materials should already be transferred from individual authors of articles to the societies. NACSIS will compile a database of journals

that are newly issued after the contract is closed. Retrospective digitization is considered separately from the digitization of forthcoming journal issues. This is because many Japanese societies do not yet have clearly enacted copyright transfer contracts between authors and societies, and the contracts for retrospective conversion tend to be complicated and time-consuming.

The tentative scheme includes two different kinds of user profile. The first is an individual user model, the second an institutional use model. We will discuss the advantages and disadvantages of both models in terms of dissemination of academic information.

7 Two Models for Use

7.1 Individual User Model

Several charging models are observed in Internet-based information services at present. One model is a service free of charges. When charges are applied, some services are provided on a fixed monthly rate with user access control. In the case of NACSIS-ELS, we will adopt a pay-per-use policy with user access control.

Thus, a user account and a password will be given to each registered user and he or she will pay a copyright charge according to the amount of usage. This is identical with conventional commercial information retrieval services. The collected charge will be distributed among academic societies in proportion to the usage of their materials.

According to the tentative agreement, societies define unit charges separately for page displays and page printouts; they can differentiate charges depending on journal titles and page attributes such as content pages, article pages and announcement pages.

As well, societies can offer discount rates for members. They can also define the delay in information availability on the Internet after print publication. Those societies concerned about losing subscriptions can set this delay for a longer period, for instance, six months.

This more complicated system has been adopted instead of a common page charge for every journal in order to provide an incentive for societies to participate in this project. For societies concerned about losing members, they can set the page access free of charge for members; those societies that generally fear the negative effects of online journals can set page charges prohibitively high. This will discourage users from using those societies' online journals.

Although such measures for price setting may confuse users in browsing pages, they will alleviate the

concerns of societies caused by the unsettling transition towards electronic journals. Unfortunately we do not yet have fixed levels of access charges for scientific information. It takes time to reach price equilibrium for online journals.

As online journals become more commonplace, societies may consider that they can earn more income through such services and they may set lower charges, more acceptable for general users. NACSIS considers that the early launching of journal digitization is most important in providing useful databases for researchers.

Some societies, however, do not have such concerns and simply regard NACSIS-ELS as a new way to boost society activities and communication among members. In such cases, societies will have no hesitation in setting the charges at a low and reasonable rate.

In the near future, electronic cash transactions will be employed with copyright protection capability embedded in the image data. This will enable us to make the charging procedure even simpler.

7.2 Problems of Individual User Model

Since the individual user model described above was based on conventional information retrieval services, total payments by users tend to be higher than in the case of using print publications at libraries in universities and research institutions.

The problem is one of who pays for the information usage. Issuing accounts for individual users only will not expand the NACSIS-ELS service. University libraries cannot afford to pay for all usage by students and patrons, because the charge tends to be high and it is difficult to estimate the necessary total. Therefore, the individual user model will apply only for a limited number of researchers who have sufficient financial resources to pay for online journals.

7.3 Institutional Use Model

The institutional use model was created following the conventional journal subscription model. Institutions pay an annual fixed charge for journal use. Societies define the rate according to the nature of the institution and the number of potential users. People attached to the institutions can access the subscribed journals freely. For non-subscribed journals they have to pay charges according to the individual user model.

The advantages of the institutional use model are:

- (1) institutions can estimate the necessary annual budget for information usage and can select journals to which they wish to subscribe;

- (2) societies can estimate the annual income from institutional users, thus making their financial state more stable; and
- (3) NACSIS can reduce the workload necessary to calculate usage and to distribute the charges among societies.

NACSIS is continuing discussions on this model with academic societies and the institutional use model will be incorporated into the NACSIS-ELS charging tariff in the near future.

8 Other Issues

8.1 Service Schedule

Until copyright charges for electronic journals are defined, the service is provided free of charge. Participating societies have accepted this proposal by NACSIS favorably.

In 1998, the individual user model charging system will be introduced. We will implement further details of the institutional use model by analyzing the usage profile and experience of the NACSIS-ELS service.

8.2 Society Activities over the Internet

Through discussion with the societies, we have found that the real problem is what role academic societies will play in the digital age. Some societies already permit members to provide their papers on their personal web sites. Strictly speaking, this may violate the copyright of the societies.

Some societies may alter and/or intensify their character as scientific publishers; others may make more efforts to strengthen communication among members in various media and to expand their publicity and educational activities through a variety of networks.

NACSIS will monitor such metamorphoses of academic societies and update the characteristics of the NACSIS-ELS service.

8.3 Technical Issues

Although password control and encryption of document images are employed in NACSIS-ELS, further complicated measures for ensuring security, such as embedding an electronic watermark into images, are not applied.

This technology may become necessary for a more secure service of copyrighted materials, but for the time being we are not introducing further measures for this purpose. It is unclear how people will accept this kind of technology and we are prepared to employ more secure capability into NACSIS-ELS if increased public acceptance indicates its necessity.

8.4 Diversity of Online Publications

Recently, several publishers have begun to provide electronic materials, and commercial servers are already working with a small number of journal titles. Too many servers with small collections would not make it easier for users to have access to a wide variety of materials. This would be so because users need to contract with numbers of service agents, and system usages tend to be different.

To establish globally distributed digital libraries, in which people can search for and browse all kinds of materials according to their interests, we have to consider equipping the NACSIS-ELS system with the following functionalities:

- (1) an archiving function in different sites;
- (2) an appropriate control function for avoiding access concentration;
- (3) a function to avoid modification of information;
- (4) a function to avoid illegal modification of information; and
- (5) uniform and common access functions to distributed servers.

In parallel with these technical developments, a system to control intellectual property rights will be required to establish economically viable distributed digital libraries that are also acceptable from a user's point of view.

9 Concluding Remarks

NACSIS-ELS can be categorized as an online document delivery system integrated with bibliographic databases, specially designed for scholarly publications. The full-fledged service of NACSIS-ELS, including copyright charge collection, is scheduled to start in October, will expedite the dissemination of scholarly information and facilitate easy access to Japanese scholarly publications, in particular, academic journals written in English for overseas scientists.

Although we give a higher priority to the digitization of current issues for the time being, retrospective digitization is also considered. Several titles have been already converted in digital form from their first issues.

The latest information concerning the NACSIS-ELS project can be accessed on WWW with URL: <http://www.nacsis.ac.jp/>.

WORKSHOP

IJWDL'98

International Joint Workshop on Digital Libraries 1998

7-9 September 1998

Asian Institute of Technology

Bangkok, Thailand

A Digital Library for K-12 Educators: the PEN-DOR project

Myron Bright, Karen Fullerton, Jane Greenberg, Maureen McClure,
Edie Rasmussen and Darin Stewart
University of Pittsburgh, Pittsburgh, PA, USA

Abstract

PEN-DOR (the Pennsylvania Education Digital Object Repository) is a digital library designed to provide access to the collective experience of teachers, students and administrators in public schools in building lesson plans and using curriculum materials. Using the WWW as a platform, PEN-DOR incorporates current research in digital libraries to provide K-12 educators with access to multimedia resources and tools to create new lesson plans and presentations, and to modify existing ones. The project is based on a distributed, object-oriented database architecture which supports the description and cataloguing of multimedia objects, and their use in support of teaching. The PEN-DOR project has elected to base its work on the GEM (Gateway to Educational Materials) metadata standard developed as part of the GEM union catalog project. Content for the database is solicited from government agencies and project partners, as well as from participating teachers. Once incorporated in the repository, materials can be organized in frameworks which form the basis for lessons, tutorials and presentations. As frameworks are developed, used, critiqued and modified, they will form a community memory of past experience. Critical issues include considerations of copyright, usability and training for a geographically scattered user community. Supported by the state's Link-to-Learn program, the system will function as a resource for educators throughout Pennsylvania.

Recent developments in scholarly publishing and their impact on libraries

Fytton Rowland

*Department of Information and Library Studies
Loughborough University, UK*

Introduction

What is scholarly publishing? I would define scholarly publications as published materials, in whatever medium, that are not primarily pedagogic in their purpose, and the content of which assumes a level of background knowledge of their subject such that only a university graduate in that subject could reasonably be expected to understand them fully. Such publications may be relatively long, in which case in their printed forms we would refer to them as books, or they may be relatively brief in which case we would call them articles or papers. In their printed versions such articles are usually, but not invariably, grouped together into issues of journals, and these issues usually, but not invariably, come out at regular stated intervals and therefore can be called "periodicals" as well as "serials". Very commonly, but not invariably, an academic publication represents the first full report of the results of a piece of research, and can thus be accorded another title, that of a "primary publication". But scholarly publications can also be of a review nature, bringing together lots of pieces of published research, summarising them, analysing them, and putting them into a structure. Such documents -- which again can be of a book form or of an article form -- are referred to as tertiary publications. In most fields, scholarly books are usually tertiary; primary publication in book form is nowadays unusual except in some fields in the humanities and perhaps the social sciences.

Those of us who study scholarly communication tend to concentrate on the journal, rather than the book, perhaps because a lot of us came out of the sciences where journal publication is more important than book publication. Both the academic book and the scholarly journal have been amazingly stable media. Despite advances in the technologies by which they are produced and manufactured, the products themselves have remained remarkably similar in character and appearance for centuries. But in the 1990s there have been rapid changes, at least in expectation if not yet in reality, in academic publication, and the brief term that can be used to describe these changes is "electronic publishing". And as you no doubt have been expecting, that is what I am principally going to talk about today.

It should be noted that traditional scholarly publications come from a variety of types of publisher. In the science, technology and medicine (STM) area, in particular, for-profit publishers are very active since the publications tend to be very profitable, owing to their attraction to practitioners and to large industrial companies as well as to academic libraries. As "publish or perish" has spread, demand for outlets for academics in other fields has grown, so the interests of these publishers have broadened to take in the social sciences and humanities too. There have been many takeovers and mergers so that a fairly small number of large international

companies now dominate the STM publishing scene. But there have always been high-quality publications produced by not-for-profit organisations, among them university presses and learned societies. Some of the smaller societies publish through a commercial publisher, taking advantage of the marketing and distribution systems of the larger organisation. Other societies, even including some quite small ones, publish themselves. There are an enormous number of very small non-profit publishers of academic material, and this presents a problem when technological innovation is required; these small organisations lack both the financial capital and the knowledge to convert their journals to electronic form. It also needs to be noted that, while these publishers are "not-for-profit", they also need to be "not-for-loss" unless they have access to some form of financial subsidy. National journals in small countries sometimes do enjoy subsidy, for example. But in general, at least a quasi-commercial attitude is required in order that the non-profit organisation can continue to publish their journals.

The purposes of the scholarly article

However, before we launch into the main part of the talk, there is one point that needs to be thought about. We will not think clearly about the future of academic publishing unless we are clear about what it is for. Many scholars, including in the UK Jerry Ravetz (Ravetz, 1973) of the Open University, John Ziman (Ziman, 1968) of the University of Bristol, and my own distinguished Loughborough colleague Jack Meadows (Meadows, 1980), have noted that the purposes of the scholarly article are in fact fourfold.

- First, and most obviously, the dissemination of information -- the

function of any publication that will most readily spring to the mind of any librarian

- Second, the preservation and archiving of that information -- ensuring that it remains available for many years after its publication
- Third, quality control -- ensuring that the information is correct and reliable
- Finally, recognition and priority for the authors -- evidence that these people did this work and had these ideas, and did so first

Academics themselves almost invariably, in my observation, tend to think of the fourth function first -- if you ask academics their opinion of journals, they will always think first in terms of "Where do I publish?" rather than "What do I read?" Some commentators are rather scathing about this tendency -- criticising the "publish or perish" mentality and maintaining that most papers are read by nobody except the authors, the editors and the referees -- and not even always by them! I prefer to take a neutral stand -- as an information scientist, I make scientific observations of the system as it is, rather than value judgements about whether it is a good or bad thing that it is so. Any proposals for change in the system, or predictions about what might happen in the future, need to take note of all the functions. As my colleague Cliff McKnight (McKnight, 1994) has observed, electronic journals must at least fulfil all of the functions of the printed journals; preferably they should do more, but they must not do less.

The electronic journal

First a word about the electronic book

The main thrust of my talk will be the electronic journal, but first a very brief word about the electronic book. There are, of course, electronic books; for example the Encarta Encyclopedia on CD-ROM is hugely successful, and electronic multimedia versions of classic children's books on CD-ROM are also selling well. But there is no detectable trend towards publication of the academic monograph in electronic form, although many publishers have web sites advertising their books. Monographs -- despite that word -- are often multiauthored, with different experts contributing different chapters, but the book has an overall structure and theme decided upon by its academic editor. One kind of book that is likely to become electronic, however, is the conference proceedings volume. Although these have traditionally been published as books, they are of course more akin to journals in that they contain many different papers from different authors describing different pieces of research. They have a general theme, of course, but unlike a multiauthored monograph they do not have a strong directing editor. Many conference organisers are now making their papers available on the World Wide Web, before or after the conference or both, and in some case this form of publication is replacing print publication of a proceedings book. This is a case where the ease of mounting material on the WWW, even by relative amateurs, is very helpful in achieving prompt and cheap publication. In many cases conference proceedings are not really a commercial proposition for publishers, either.

Moving on to the main theme, electronic journals, I will now deal with their development historically.

Early experiments

The Information Explosion -- now rechristened Information Overload -- has been talked of since the 1950s, if not before. Once the computer was invented, scientists naturally thought of using it to help cope with their problem of too much information. Electronic versions of the major scientific abstracts and indexes journals have existed since the end of the 1960s, and online information retrieval services operating in real time, based on those secondary databases, have been operated by Dialog and its competitors since the mid-1970s. Given the obvious advantages of searchability and space-saving, to name only two, it is unsurprising that information specialists then began to turn their minds to the possibility of producing electronic equivalents of the primary journal as well. In about 1976-77 the first couple of experiments in online primary journals began, and John Senders (Senders, 1977) made his famous comment "I have seen the future and it doesn't work". He was, of course, right -- and it still doesn't work very well. The difficulties in these early years were essentially as follows.

- The only network generally available was the voice telephone network, which then consisted of copper wire and Strowger electromechanical switches. The charges for long-distance calls were high and the quality of lines unpredictable.
- Few scholars had ready access to computers, and fewer still could connect them in any way to the network.
- The early microcomputers -- Apple 2s and Commodore Pets among others -- were incompatible with mainframes and with each other. The software could not easily exchange files between different types of machines,

which were not designed for mixed-platform networking

- Neither the hardware nor the software could, in general, cope with anything other than alphanumeric characters: graphics of any sort were very hard to display
- Neither publishers nor libraries were equipped to handle electronic publications. Some journals were still printed using hot metal; even where computer typesetting was already in use, every proprietary system used its own codes and all were mutually incompatible.

The BLEND (Shackel, 1991) and Quartet (Tuck *et al.*, 1990) experiments, with which Loughborough university was heavily involved, were major British electronic publishing experiments. Despite these problems, much was learned by the early experimenters. In particular they investigated the possibilities for interactive journals -- that is, where the readers could respond to, annotate and criticise the published papers. projects of this era. This is a theme to which I will return at the very end of my talk.

The "free" journals

As the Internet began to develop and become widely available, and high-specification PCs came down in price, at the beginning of the 1990s some scholars began to advocate a completely new method of scholarly publication. Arguing that most scholarly journals were of interest to only a few people throughout the world, they felt that the commercial system of publication of scholarly primary articles was inappropriate; copies were sold at high prices to university libraries, when it was feasible to distribute the same material direct from academic authors to

academic readers via the Internet at no direct cost. (There were, of course, indirect costs such as the provision to academics of PCs, network connections and national and international academic networks, but it was assumed that this was happening anyway.) Advocates of this approach, notably Stevan Harnad (Harnad, 1996), then at Princeton but now at Southampton University in the UK, and Jean-Claude Guédon of l'Université de Montréal in Canada, said that each journal could be edited by its academic editor (such as Harnad or Guédon themselves), mounted on the Internet using generally available and user-friendly software, and downloaded for reading or printing by interested readers anywhere in the world without the intervention of either a publisher or a library. It is indeed the case that a substantial number of new journals of this kind have been established. Originally, in 1990-91, they tended to publicise themselves initially on suitable e-mail discussion groups; they then set up their own discussion group for the journal itself, to which interested readers were invited to subscribe; this e-mail list then received the contents pages and abstracts of each issue of the journal, and the full articles including graphics could be downloaded using anonymous ftp. This was not especially user friendly, and when first gopher and then the World Wide Web came along the embryonic electronic journals quickly switched to these friendlier packages.

New commercial journals

Over the years since the emergence of the World Wide Web -- since say 1992/93 -- a number of new electronic-only journals have been established by the major scholarly journal publishers, such as Oxford University Press and Springer Verlag. It has to be said, however, that there are not a large number of these. The

reasons might be: (1) It is difficult for any new journal -- paper or electronic, published by a commercial or a non-profit publisher, to get established; unless there is an obvious gap in the market, it is hard to compete with established titles. (2) Charging methods for purely electronic journals are only just emerging, and it is difficult for the publishing to get pricing structures and pricing levels right. In our own field, I might mention the *Journal of Digital Information, JoDI*, edited by my Loughborough colleague Professor Cliff McKnight jointly with Professor Wendy Hall of Southampton University, and published in association with Oxford University Press. I think the jury is still out on the question of whether such journals can pay their way: even traditional printed journals took up to five years to break even from their initial foundation, so it would be surprising if ones founded in the new medium could be in the black yet.

Dual journals

Where there has been the most rapid growth, however, is in well-known, well-established journals from the traditional academic publishers, in both the for-profit and the not-for-profit sectors, being made available in electronic form in parallel to their continued publication in printed form. Most of the significant publishers in the scholarly journals business have by now made their journals electronically available, a few from 1995, more from 1996, many more from 1997, and some relative laggards from 1998. I will discuss the question of payment for these journals later.

Details vary, but typically a publisher will make available the full text and graphics of all the papers -- but not always all the ancillary matter -- on the World Wide Web. The entry screens to the publisher's site will typically be freely available, as

often will be the titles and abstracts of the individual papers. Sometime a few papers will also be available free of charge as "tasters", but the general run of papers in full text will be available only to paying customers, and access is protected by either a password system or checking of IP addresses of the customer's machine. Paying mechanisms also vary and are discussed later. The most usual format for the presentation of the full texts is PDF (publications distribution format) which is derived from PostScript and handled by the Adobe Acrobat software. This preserves the page layout of the printed publication (so-called page integrity) but allows searching of the full text. Contents pages and free abstracts, on the other hand, are usually in HTML, which is quicker to download than PDF but does not preserve the appearance of the printed original -- the appearance will depend on the user's browser settings. Some, but not many, journals provide an HTML alternative for the full texts, and some publishers have opted for other page-integrity systems, notably RealPage from the software house CatchWord.

New features

A major argument for electronic journals -- whether free or for fee -- is that an electronic publication can do things that a printed one cannot. First among these considerations is full-text searchability. Secondly, a structure of hypertext links, both within and between documents, can be put in place, and work at Southampton University, for example, is looking into the possibilities of separating out the link structures from the actual content itself. Thirdly, multimedia content is possible -- an article can contain video, sound and animations as well as text and still pictures. Fourthly, fuller data can be included which it would be uneconomic to print. Finally, articles can contain runtime

versions of software, allowing manipulation of the data by the user, or the user can extract data from published sources and manipulate them using software of the user's choosing.

Informal communication

There is a body of opinion, whom one might call "Internet libertarians", who believe that formal publication of scholarly material is no longer necessary in the Internet era. Everyone can, and should, publish their own views on their own World Wide Web site, or within suitable e-mail discussion lists, and nobody should be allowed to censor what they say. The refereeing procedures applied by publishers of traditional academic books and journals are an anachronism resulting from the previously unavoidable costs of printing; because printing and distribution of print publications are expensive, space in them has to be rationed. Space on the Internet is effectively infinite, so everybody can be allowed to say as much as they want to without restriction.

Obviously this is a caricature of the viewpoint. However, it does raise a serious point. Did traditional scholarly publishing silence heterodox views? Does electronic publication offer a new and valuable academic freedom? Can the Internet provide help for the "little person" against the big battalions, of governments or of large corporations? I do think that it is important to recognise the enormous value of improved informal communication between scholars. But it is also important to know the difference between the informal discussion and formal publication, and to value both.

E-mail discussion lists and newsgroups

Even before the emergence of large numbers of formal scholarly journals published over the Internet, academics had been quick to adopt the Internet as a tool for their work. Electronic mail has transformed academic co-operation: it is now possible cheaply and easily to communicate with colleagues throughout the world on a close to real-time basis. E-mail discussion lists extend the one-to-one e-mail concept to communication within a group of people sharing an interest. Ideas can be tossed around, data distributed, drafts of articles circulated among dispersed authors, and so on, with convenience. More formal "computer supported co-operative work" (CSCW) using software like Lotus Notes, or more sophisticated equipment such as whiteboards on which all parties can write, each at their own terminal, and cinecameras attached to machines so that the participants can see and talk to each other at the same time as they work on their PCs, are leading to what has been termed "collaboratories". There is little doubt that these advances have improved academic research productivity. It is, however, important to distinguish between this kind of informal discussion and publication. As I argued in an article last year in *Ariadne* (Rowland, 1997), we must remember that many real-world decisions depend on the availability of published data of known quality and reliability. This is perhaps more true in the physical sciences than in the social sciences. But I would not like to see government making significant changes in social policy based purely on rumour and Internet chat. One would like to see rigorous social-science research being used to underpin policy. This means that the published results of such research have to be subject to quality control and then made available, in the long term, in ways that prevent their subsequently being altered, by their own author or anyone else. This is not to say,

of course, that online comment and discussion of the published results would not be an enhancement of a journal. Indeed, as we have seen, this possibility was one of the first features of electronic journals to be investigated in the early experiments such as BLEND (Shackel, 1991).

Preprint exchanges

Long before the emergence of the computer as an information-handling tool, scholars had circulated drafts of their papers to each other for comment, and after the arrival of the photocopier, it became possible for authors to send out substantial numbers of copies of their draft papers to professional colleagues. However, as this occurred in advance of the acceptance of papers by a journal, the quality control element was missing -- the paper might not have been published in the same form, or at all. Thus they were known as preprints. Efforts were made from time to time to systematise the preprint distribution process, but many scholars opposed this on the grounds that it did circumvent the quality control procedure.

Recently, with the general availability in developed countries of Internet connections, and word-processing software enabling almost all academic authors to prepare their articles in machine-readable form, the pre-print exchanges have become electronic. In high-energy physics, in particular, the exchange organised by the Los Alamos laboratory in the USA has become the major current-awareness tool of the discipline (Ginsparg, 1994). Papers can be read on the Los Alamos server many months before the appearance of the journal, and thus the journals themselves are tending to be purely archival devices.

This may militate against good sales of physics journals in electronic form.

WWW sites

It is a commonplace among those of us who teach undergraduates at the end of the 1990s that it become increasingly difficult to persuade them to look anywhere other than the World Wide Web for all the information that they require! Even Information and Library Studies students prefer the World Wide Web to the library. Nevertheless, it has to be remembered that WWW sites occupy border territory between formal and informal communication. Sites of major research universities, "resource discovery" sites like SOSIG (social sciences) or OMNI (medicine) in the UK, and the sites of major scholarly publishers clearly do provide access to information of tested quality. At the other end of the scale, personal web sites of private individuals may simply reflect their own personal prejudices and enthusiasms, without any guarantee of accuracy or objectivity for the information contained

Blurring of the boundaries: is it a good thing?

I believe that both informal discussion among academics, and formal publication of research results, are necessary for the good health of the scholarly enterprise. One does not replace the other. Returning to Cliff McKnight's (1994) principle that the electronic publication must at least do all the things that the printed one does, and Jack Meadows' (1980) list of the functions of the scholarly journal, we see that electronic scholarly journals must provide for the dissemination, archiving, quality control and recognition functions of the print journal. In principle they can do all these things, but only if a proper degree of formality exists. Electronic discussion,

and the free and uncensored expression of opinion, are good things in their own right. They do not replace, however, the publication of tested, authenticated data, whether in the physical or the social sciences. The humanities may be a rather different matter; but there too, presumably, academic authors would want to ensure that their own views are accurately presented and not subject to subsequent tampering or misrepresentation by others. (This is the concept of moral rights, enshrined in European Union copyright legislation.)

One thing that does seem clear is that academics are using all the informal channels themselves, without the intervention of librarians. Some may need training in the use of the IT and Internet facilities involved, and this training may be given by their campus Computing Service or by the library, depending on local circumstances. Once they are trained, though, they will in all likelihood do their own netsurfing.

Financial considerations

Means of payment

Returning, then, to the formal electronic journals, what is the involvement of the library with them?

In the case of the free ones, there does not inevitably have to be any library involvement: users can access them directly over the Internet without formality. However, not all potential users necessarily know how, and some may not even know that they exist -- one important role of commercial publishers is marketing, and for free journals this activity may not be done, or not done well. So the library can tell its users that the journal exists, and smooth their way towards using it, most likely by including

a link to it in the library's own web pages. At Loughborough, the library has a web page called "Electronic Journals" with a link from the library's home page. The electronic journals page in turn gives access to four more pages: one for free journals, one for sample free issues of commercial journals to which the library does not subscribe, one for the Blackwell's Navigator service (of which more later), and one for commercial publishers' services to which the library does subscribe, whether through the Pilot Site Licence Initiative or otherwise. The free journals page in turn lists by name those free journals which, in the academic librarians' judgement, are not only relevant to subjects studied at Loughborough, but also of an adequate academic standard to be worthy of recommendation. There are then links directly from the name of a journal to its own home page. Other universities do something similar. In some cases, however, academic departments may do the job themselves rather than leaving it to the library. The department of chemistry at Cambridge University in the UK, for example, has an excellent chemistry resources site, with links to chemical electronic journals among many other things.

The complicated work for libraries arises if paid-for electronic journals are provided. To buy print journals libraries mostly use subscription agents. These companies reduce the administrative burden of dealing with multiple publishers, at a relatively modest charge, by acting as a middleman (or hub, in airline parlance) between lots of libraries and lots of publishers. Now that all the major ones are fully computerised their services are very efficient.

With electronic journals the situation is much less tidy. Publishers clearly have not realised

what an enormous burden of unnecessary and unproductive work it gives to libraries if they have to deal with every publisher separately -- especially if publishers' systems are incompatible with one another. (And let us not forget that publishers, accused of overpricing, often suggest that libraries should spend more on purchases and less on staff.)

The payment options offered are generally these -- though not all publisher necessarily offer all the options.

Annual subscription to individual journals, as with the print products. Within this option, there may be the choice of buying the print and the electronic version, the print only, or the electronic only. But not always -- one publisher at least is now making everyone pay for the electronic version -- at an extra charge, naturally -- whether they want it or not; the option of buying only the print has gone. Another -- a very reputable American learned society -- lets you buy electronic alone, but the print version is now an add-on, naturally at an extra charge, and cannot now be bought alone. Many will not let you buy the electronic without the print.

Subscription to all of the publishers' journals in electronic form. If this option is taken, there is a saving over the total cost of buying them separately. But of course publishers (other than learned society publishers) do not cover just one discipline nor does any one publisher usually produce all the journals in any one discipline. The publisher's stable of journals is a rag-bag of journals across miscellaneous fields. It is unlikely any library -- except the very biggest ones -- would actually want all of the journals from any one publisher.

Terms of different publishers' licence agreements vary: different publishers will

allow their customers to do different things and forbid different things. This is very difficult for librarians, as their users naturally do not understand why they can do something with one of their favourite journals but are forbidden to do the same thing with another from a different publisher. Publishers are naturally concerned about illicit copies of their content escaping and thus undermining their revenue, and so try to restrict access within universities. This can be done, for example, by requiring the library to issue all users with individual passwords, or by only allowing access from machines with the appropriate IP addresses that associate them with a subscribing university. Some have tried to insist on access from machines actually physically located in the library, but this has been abandoned because it negates the main advantage of electronic publishing -- access at the user's own desk. Some find it difficult to cope with a type of user who use many different machines all over the campus at different times, unpredictably. This type of user is called a student, using open-access PC labs. Some have tried to limit the number of simultaneous accesses by users within the same university. In one case they are trying to limit this to one user to any one title at a time -- an access far inferior to that provided with print journals, where after all one reader can be consulting volume 27 while another reads volume 34! This seems to me an offer that customers are bound to refuse.

National or regional site licensing or consortium purchasing. In response to the kind of difficulties that I have just been describing, another option is being examined in many parts of the world: national or regional site licensing or consortium purchasing. In the USA groups of universities are banding together in consortia to gain bargaining power with publishers, and Elsevier, indeed, are

insisting that they do so as they do not want the administrative burden of dealing with 2,000 US universities any more than any university wants the burden of dealing with 30,000 publishers.

In the UK the Pilot Site Licence Initiative (PSLI Evaluation Team, 1997) has been running for a couple of years now with four publishers: Academic Press, Institute of Physics Publishing, Blackwell's and Blackwell's Science. These publishers agreed to supply their journals -- print and electronic -- to all UK higher educational institutions in return for a payment made to them centrally by the Funding Councils. In Academic Press's case, no further payment was requested from the universities -- i.e. they received a 100% discount off the normal prices -- while the other three publishers provided reduced subscriptions (i.e. discounts of less than 100%). This project, seen as a pilot scheme as its name suggests, was controversial because it was compulsory -- that is, the funds were top-sliced by the Funding Councils so that universities were indirectly paying for these journals whether they wanted them or not.

Now the National Electronic Site Licence Initiative, NESLI, is offering a voluntary alternative (Friedgood, 1998). The Funding Councils are providing pump-priming money for three years to set up what might become a self-funding scheme eventually. This time it deals only with electronic subscriptions -- the printed versions are not involved at all. This in itself is proving difficult as many publishers do not wish to decouple print and electronic subscriptions. Two managing agents have been appointed -- Swets and Zeitlinger, the subscription agent, will look after negotiations with universities and publishers (acting as the university libraries' agent) and administrative aspects, while Manchester

Computing Centre at Manchester University will deal with technical aspects, mounting the NESLI homepages, maintaining the hypertext links with all the publishers, providing search facilities that will work across all the publishers, and so on. NESLI's work has only just started but it is clear that negotiations with publishers will be very tough. A Standard Licence has been drawn up, but individual publishers may negotiate different terms.

Pay-per-view or pay-per-print. This option is being provided by many publishers in recognition of the fact that private individuals, or indeed small companies or institutions, may wish to have occasional access to a journal to which they would not subscribe. Thus the option of paying just for the one item that you happen to want is also made available. Some American information professionals call this pay-per-drink -- what you do when you buy a single whisky in a bar rather than buying a bottle of whisky in a liquor store. This option of course blurs the distinction between publishing and document delivery services like those of British Library Document Supply Centre in Boston Spa, UK, and other commercial suppliers. There are technical problems because of the difficulty of collecting micropayments -- that is, the administrative cost of collecting a small amount of money far exceeds the amount paid. However, information technology will probably provide a solution here, as many software houses are developing Internet micropayment systems to enable people to pay small sums by credit card at a very low administrative overhead cost. Security is also a problem, as many people do not wish to key their credit card number into the Internet which is known to be leaky. Again, software houses are working on appropriate systems for secure Internet payments using encryption technology.

At one time it was suggested that this method of payment might completely replace the subscription methods. It is now recognised that such a scenario would not provide either publisher or library with stability. The publisher needs to know that there is enough firm revenue coming in to cover the "first-copy cost" -- that is, the cost of putting the information product together in the first place. The subscription basis provides for that, and also provides publishers with a positive cash flow -- subscriptions are paid at the beginning of the year while costs are incurred throughout the year. It also provides libraries, which are usually not revenue-earning departments, with stable and budgetable costs; the library will know in advance how much money its institution will give it for the year, and therefore needs to know how much it is going to have to spend on journals, in order to budget. Pay-per-view does not provide such predictability. However, pay-per-view as an alternative to subscriptions is attractive to publishers, since it offers the possibility of new revenue sources -- individuals or institutions that otherwise would not have purchased the title at all now provide it with some revenue, even if only a small amount. A large number of small payments that the publisher might not otherwise have received at all could be an important extra revenue source, which in turn could either increase the publishers' profit or allow them to set subscription rates lower than they would otherwise have had to.

Who needs libraries?

It has been suggested by some people that libraries would be unnecessary because everyone could simply access the free electronic journals on the Internet for themselves. This is true, but experience at Loughborough in various users studies

(Rowland *et al.*, 1996; Meadows *et al.*, 1997; Woodward *et al.*, 1998b) suggests that they don't use them, mainly because either they do not know that they exist -- remember marketing is a key role of publishers -- or because they don't know how to use them -- remember user education is a key role of libraries.

But it is clear that the swing away from commercial scholarly publishing to the free variety will be slow and probably will not be complete. Indeed, as has been noted, even the current free ones may have to institute modest charges when their subsidies cease. And if there is any charge at all, then there will be administration needed to deal with the payment and collection of the money. In the UK, at least, it is unlikely in the extreme that academics will be willing to pay for journals out of their private pockets. Students too can reasonably expect that access to scholarly journals is one of the things they are paying their tuition fees for. Thus some part of the university structure will have to deal with the negotiation with publishers, with the making of decisions on which products to buy and which not to buy, and with ensuring that the access which they have paid for is actually reliably provided. The obvious part of the university structure to perform these tasks is the library. And of course library staff are the obvious people to advise staff and students on the choice of information resources to use -- as they always have done -- and to teach them how to access them. This too was always necessary -- few students arrive at university with an intimate knowledge of the Dewey Decimal Classification! And the user-training role becomes greater as a large variety of electronic products are provided. They should, no doubt, be user-friendly and transparent, but they are not; and as technological obsolescence grows ever faster, there is little chance that any

users unaided will be able to keep up with the avalanche of change. It will always fall to an information expert to keep up with the changes and organise ongoing training programmes for users. And who is better qualified to do this than the library staff?

Even if the research support role of the library becomes wholly or almost wholly electronic, it is unlikely that the teaching support role will. Thus universities will continue to have libraries, and some -- albeit a declining proportion -- of the information resources within them will be in print. It makes sense for one focus to provide access to all information resources regardless of medium. It would be uneconomic and inefficient for every department or faculty to try to organise access to electronic information resources separately -- not least because several departments might subscribe to the same title while another of minority interest to several departments was purchased by none of them.

Another issue that has seen much debate lately has been the archiving and preservation of electronic journals. If I buy an annual subscription to a printed journal, and then cancel the subscription, I keep the issues for that year: the physical journals are my property, though their content is not my intellectual property and I am constrained by the law as to how much copying I can allow. But I could lend the issues that I had bought to someone else, sell them second-hand, give them away or bequeath them. If I buy a year's subscription to an online journal and then don't renew it, on the other hand, I have nothing for my money. At the end of the year my access is withdrawn.

Publishers have never provided archival access to their publications; once they are out of print and no copies are left in the

warehouse, the only place to get a copy is from a library. It seems reasonable to suggest that libraries should provide archival access to electronic journals too. In the electronic medium, one technically copies an item when one prints it out or downloads it. When I put a printed journal that I have bought on to a library shelf, I do not copy it. Thus technically doing anything at all with an electronic journal, even when you have subscribed to it, entails copying it and is therefore restricted by copyright law; unless your agreement with the publisher says you can do something, you can't. Unless publishers are reasonable about allowing libraries (or somebody!) to retain copies of their journals indefinitely, there is a danger that electronic-only journals might disappear altogether. There is no financial incentive, no business reason, why the publisher should retain material that will generate no more revenue. So why are some publishers trying to stop other bodies from keeping their material for the long term? As yet this is not a major problem because electronic-only journals, not printed at all, are in a small minority and all the major journals still appear in print. But that situation will not last much longer. Some publishers issue a retrospective CD-ROM at the end of the year, or even a retrospective printed volume, as part of a library subscription to their electronic-only journals -- these seem to me to be far-sighted measures. The issue of how to preserve journals whose publishers are not so far-sighted is unresolved. Personally I would support an extension of legal deposit to electronic publications, so that the British Library and equivalent legal deposit libraries in other countries are permitted to download and retain a copy of every electronic publication, to ensure its permanent preservation.

So, speaking as an academic who is not a librarian, I affirm that we will need

libraries and librarians in the brave new electronic future!

The future of the scholarly journal

If I had been delivering this paper a couple of years ago, I would have said that the commercial publishers of academic journals were likely to find themselves in difficulties -- of their own making, owing to their refusal to reconsider their excessive pricing. I believed that the various initiatives being taken by academics, university librarians and university administrations to bring into being an alternative non-commercial electronic publication system might bear fruit. The driving force, of course, was the desperate shortage of funds for academic libraries, but the well-publicised profit figures of the Reed Elsevier group from the Elsevier Science division, and the very high price increases imposed by certain smaller commercial publishers, also contributed to the feeling that the commercial publishers were heading towards the edge of a cliff.

Now I am not so sure. Although new non-commercial electronic journals continue to appear, many of them have only start-up funding and they are all having to look towards cost-recovery business plans.. In our own field *Ariadne* is a successful e-journal -- my paper last year (Rowland, 1997) in that journal has generated more comments back to me than anything else I've ever written, suggesting that people actually read *Ariadne*! But with the end of their funding they will have to start charging.

Meanwhile the mainstream scholarly publishers, both for-profit and not-for-profit, have been very active. For higher education libraries at least, the most notable activity has been, of course, the Pilot Site Licence Initiative (PSLI Evaluation Team,

1997) and its successor the National Electronic Site Licence Initiative, NESLI (Friedgood, 1998). The implicit assumption behind these activities of the UK Higher Education Funding Councils (and presumably therefore also the assumption of the UK Government) is that the commercial nexus is here to stay in the scholarly journals field. I am bound to say that I find this disappointing. I had hoped that the higher education institutions collectively -- might have been able to agree on some collective action to fight overpricing. We could have organised a monopsony - that is, a situation where there is only one customer! But the natural competitive tendencies of universities seems to have militated against this, and the publishers, despite all the complaints, are clearly still confident that they can continue to compensate for lost subscriptions by raising prices to their remaining customers by figures well above inflation. I still think they will fall off the cliff one day, but the lack of will to organise an alternative has postponed that day into the 21st century.

Clearly, no-one can go on raising prices indefinitely when their chief market has declining purchasing power. Charging "what the market will bear" is fair enough in a market economy, but charging beyond what the market will bear is foolish in the long run, even if profitable in the short run. Publishers have, of course, incurred considerable investment costs in tooling up for electronic publishing, and quite rightly point out that during the era of dual publishing they still incur the direct costs of print publication: paper, printing and distribution bills. If and when publication becomes purely electronic these costs will cease -- at least for the publisher, though of course paper and laser-printing have to be paid for by the end-user of electronic publications, if they print articles out. Meanwhile publishers do have a case that

their costs have actually risen. In trying to recoup from their existing, paper journal, customer base the costs of developing the electronic alternative -- rather than by taking the investment funds out of their existing profits -- they risk destabilising the entire scholarly publishing chain.

There is another reason why I am less optimistic about the e-journal future than I was a couple of years ago. Many university libraries are now making a substantial number of electronic journals available to staff and students via campus networks, as noted earlier. Do any of you have an evidence that anyone is using them? A graduate student of mine is doing her dissertation about e-journal use in the six universities in our region of England, and while she has had success in holding extensive interviews with helpful library staff at all six universities, her attempts to find a population of users of e-journals to interview among the academics have been an almost complete failure, despite extensive appeals. And Loughborough, at least, has been among the vanguard of universities nationally in encouraging use of electronic information resources by staff and students. Publishers, and librarians, need to have regard to the needs and wishes of the real end-customer, the academic reader, as well as the views of the purchaser, the university library. I heard recently about a Scandinavian university where the librarian has decided to transfer all significant support resources for journals to the electronic version; where publishers insist on bundled subscriptions and are therefore still sending printed issues, these are placed on the shelf but nothing is done with them. This is a brave approach, but I do wonder what the academics and students at that university really think about the policy.

So where does all this leave the scholarly journal in the academic library? In a state of flux, I suppose! The free electronic journals will grow slowly in number and size, but will have to start charging at least modest prices to cover their production costs. Overt or covert subsidy will not last for ever, unless here really is a fundamental redesign of the scholarly communication system. This would have to be led by the academics themselves, though either their universities or their learned societies, and present evidence is that it is rather unlikely.

The major publishers will continue to move down the road towards dual publishing until all existing scholarly journals are available in both forms. Their content will start to diverge, as presaged by the SuperJournal project in the UK -- with multimedia content, hypertext links and supplementary data included in the electronic version. A variety of pricing options will be created by publishers, including involvement in licensing schemes like NESLI, annual subscriptions to all of a publisher's output, annual subscriptions to individual titles, and pay-per-drink. The last-named is unlikely to appeal to libraries, but its importance is that it just might produce a new source of revenue for publishers. The real answer to the journals pricing crisis is for some of the first-copy cost of journals to be paid for from somewhere other than academic libraries' acquisitions budgets. If purchases of individual articles -- currently handled through document delivery companies -- moves in the electronic era back towards the publishers themselves, they may start to get revenue from departmental or even personal pockets in addition to that derived from the library.

Another issue for libraries is the division of their acquisitions budgets between different types of product. For years there

have been complaints, especially from academics in the humanities, that the rising prices of journals, especially in the sciences, have led to libraries cutting their book purchases to afford the essential journals, and as I am sure you all know many libraries have responded to this criticism by trying to redress the balance. There is also the issue of the balance between the library as a research resource, mainly provided by journals, and the library as a teaching resource, substantially provided by books and often in multiple copies. This balance will vary from university to university depending on the research-intensiveness of the institution, but every university does have to teach. Many lecturer's reading lists today will include WWW sites, and these have the advantage that students will not find that all the copies are out on loan. Many lecturers also specify particular journal articles, too, which has led to experiments on the question of the electronic equivalent to short-loan journal article photocopy collections. One of these experiments at Loughborough, ACORN, was very successful and has demonstrated a possible way forward for this area of library support to teaching (Woodward *et al.*, 1998a). But all the indications are that the bulk of lecturers' reading lists will continue to be books, essentially textbooks. This may change slowly as more computer-assisted flexible learning comes into use in universities -- another interesting topic that is perhaps slightly outside the scope of today's talk. My point now is that the pressure on library budgets from the teaching and learning function, as contrasted to the research function, of the university, will continue to be strong.

Finally there is competition between electronic products, as between primary and secondary materials in electronic form. For years libraries have been buying relatively expensive abstracts and indexes

databases on CD-ROMs, and now some of these are being mounted on hard disks within the institution (after delivery on CD-ROM) to overcome the technical problems of CD-ROM networking. As primary publications become available in electronic form, there are on the one hand budgetary pressures as between the two types of electronic resources, and pressures from users that the two should be linked. Thus one would like to be able to conduct a search in *Chemical Abstracts* databases, say, and then link seamlessly into the full texts of the articles retrieved. The technical, financial and contractual problems for the library, however, in making such a user-friendly service available are not trivial. Every such enhancement of service is expensive, not just in terms of the subscriptions to the products, but also in terms of the high-calibre administrative and technical support staff time needed within the library.

Some of the secondary database providers are, of course, the same firms that publish primary journals. Thus Elsevier now owns not only *Excerpta Medica*, theirs for some time now, but more recently *Engineering Index* too. The American Chemical Society owns not only *Chemical Abstracts* but also an impressive stable of chemical primary journals of high repute. But this does not solve the problem of seamless linking: any secondary database, in order to be comprehensive, will have to cover the primary journals of many publishers including many very small not-for-profit ones. So the users' need for hypertext linking across the World Wide Web between the products of many competing publishing companies will remain, and this is one that is going to be hard to solve technically, financially and even legally (given the restrictions that exist in many jurisdictions on collaboration between competitors).

In an ideal world, the user, a university staff member, would be able to use a WWW browser on their own desk -- at the university or in their study at home -- to access their university's information resource (the thing we might old-fashionedly call a library). They could then, in a straightforward way without needing any degree of IT skill beyond that which is now commonplace among educated people, navigate through the various resources, primary and secondary, to which their institution subscribed or which are available free of charge, find the ones that they need, and, if they wished, print them out. (You cannot of course, print out multimedia features, but they could be downloaded to the users' own hard disk.)

If the user is a student, you will need to provide the same service from open-access PC labs throughout the campus, as well as from network points in rooms in halls of residence, and, by using the public telecomms network, from students' homes or lodgings in the town as well. The University of Wolverhampton has a fascinating project here, in co-operation with their local cable TV company to provide access to the university's system from any building in the town, for users with the appropriate authorisation.

The difficulties in a library providing this sort of service are currently immense, and most stem from the competitive behaviour of academic publishers. Thus, in the print era, every publisher was willing to sell all their journals through subscription agents in what I think we can agree was a convenient and efficient system. Now, however, the different subscription agents and some other organisations are offering rival electronic journal delivery services -- one can instance Blackwell's Electronic Journal Navigator as an example of what I

mean (without in any way endorsing that one over its rivals!). I would like to see all of the publishers allowing all of their electronic journals to be accessed via all of these services, so that the end-customer, the university library, could make its own choice as to which of these "one-stop shops" they would subscribe to, just as they do now with subscription agents. Alas, however, you cannot do that. Every "one-stop shop" has signed up a different subset of the publishers, and the largest academic publisher of them all, the Reed Elsevier group, refuses to go into any of them because they have established their own, ScienceDirect, to which they are also trying to attract other smaller publishers.

The NESLI project (Friedgood, 1998) clearly aims to help with this problem, at least so far as UK higher education institutions are concerned. If publishers can be persuaded to enter the scheme, the managing agent, Swets and Zeitlinger, will fulfil the "middleman" role that subscription agents filled in the print era, relieving the libraries of the administrative burden of dealing with each publisher individually. The Manchester Computer Centre will similarly relieve them of the technical work of ensuring good reliable links to all the publishers' sites, and provide cross-publisher searching, and linking between publishers and to secondary databases. No university is compelled to take part; they can still deal with publishers direct if they wish. No publisher is compelled to take part either, and if they do take part they can negotiate variations from the standard contract. It is not yet known what pricing deals Swets will be able to negotiate with publishers, but the more universities come into the scheme, the stronger the bargaining power Swets will have.

I am sorry that I cannot give a clearer picture of where things are going. All I

can offer you, in Churchillian terms, is plenty of blood, toil, sweat and tears, as you try to deal with this technically, legally and financially complicated situation on behalf of your users, out of your shrinking budgets, and often with declining numbers of those very staff that are needed to cope with it!

REFERENCES

Friedgood, B. (1998) The UK National Site Licensing Initiative (NESLI). *Serials*, 11(1), 37-39

Ginsparg, P. (1994) First steps towards electronic research communication. *Computers in Physics*, 8(4), 390-396, and <http://xxx.lanl.gov/blurb/>

Harnad, S. (1996) Implementing peer review on the net: Scientific quality control in scholarly electronic journals, in *Scholarly Publishing: The Electronic Frontier* (Peek, R.P. and Newby, G.B., eds). Cambridge, Mass.: MIT Press, pp. 103-118

McKnight, C. (1994) Network scholarly publishing: retrospect and prospect. In *Internet World and Document Delivery International 94*. London: Mecklermedia, pp. 86-92.

Meadows, A.J. (ed.) (1980) *Development of Science Publishing in Europe*. Amsterdam: Elsevier.

Meadows, J., Rowland, F. and Yates-Mercer, P. (1997) An online electronic journal for teaching purposes. *ALT-J, the Association for Learning Technology Journal*, 5(1), 13-18.

Pilot Site Licence Initiative Evaluation Team (1997) The UK Pilot Site Licence

Initiative: A progress report. *Serials*, 10(1), 17-20.

Ravetz, J.R. (1973) *Scientific Knowledge and its Social Problems*. Harmondsworth, UK: Penguin Books

Rowland, F. (1997) Print journals: Fit for the future? *Ariadne*, 7, 6-7; longer version at <http://www.ariadne.ac.uk/issue7/fytton/>

Rowland, F., McKnight, C., Meadows, J. and Such, P. (1996) ELVYN: The delivery of an electronic version of a journal from the publisher to libraries. *Journal of the American Society for Information Science*, 47(9), 690-700.

Senders, J.W. (1977) An online scientific journal. *Information Scientist*, 11(1), 3-9.

Shackel, B. (1991) *BLEND-9: Overview and Appraisal*. British Library Research Paper 82. London: The British Library.

Tuck, B., McKnight, C., Hayet, M. and Archer, D. (1990) *Project Quartet*. LIR Report 76. London: The British Library.

Woodward, H., Gadd, E., Kingston, P., Goodman, R. and Muir, A. (1998a) *Project ACORN (Access to Course Readings via Networks) Final Report*. Loughborough, UK: Pilkington Library, Loughborough University.

Woodward, H., Rowland, F., McKnight, C., Pritchett, C. and Meadows, J. (1998b) Café Jus: An electronic journal user study. *Journal of Digital Information*, 1(3), <http://jodi.ecs.soton.ac.uk/Articles/v01/i03/Woodward/>

Ziman, J.M. (1968) *Public Knowledge: The Social Dimension of Science*. Cambridge, UK: Cambridge University

P e s s

Retrieval System for the Microfilm Image Databases in the Media Center, Osaka City University

NAGATA Yoshikatsu, SHIBAYAMA Mamoru, KITA Katsuichi, MAEDA Harumi
Media Center, Osaka City University
3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585, JAPAN
Email: {nagata, sibayama, kita, harumi}@media.osaka-cu.ac.jp
Fax: +81-6-690-2736 (+81-6-6690-2736 from Jan. 1999)
URL address: <http://www.media.osaka-cu.ac.jp/indexe.html>

Abstract

Historical materials and other important documents are often recorded and stored on microfilm. Despite the importance of such primary materials in historical studies, however, their accessibility is limited by the need for the user to handle the microfilm in person.

To increase the accessibility of microfilm resources, a retrieval system for microfilm image databases has been implemented in the Media Center, Osaka City University. The basic concept of this retrieval system is to provide a user-friendly interface that is platform-independent.

Although the system includes mechanical devices, no serious trouble has occurred since we started operation two years ago.

In this paper, we introduce the concept and implementation of this retrieval system and related facilities in the Media Center.

1 Introduction

Primary materials are of prime importance to historical studies. A given text may exist in several versions or with various interpretations, and each material should be treated as a unique object of study.

Although historical materials should be freely available to researchers, access to rare materials is usually strictly limited to qualified and authorized researchers in order to minimize the risk of damage. For wider access, original materials are usually copied onto microfilm.

Even so, researchers who want to access these materials need to handle a microfilm reader in person. And remote users have further inconvenience in browsing the microfilm. Also, from the viewpoint of library routines, personnel are required to manage the lending service of microfilm cartridges.

The Media Center, Osaka City University, holds several rare collections, and numerous selected items are duplicated on microfilm. To facilitate and increase the use of these microfilm resources, a retrieval system for the image databases of microfilm (Microfilm image Information Retrieval System, MIRS) has been implemented. This keyword-based image retrieval system has been developed on a workstation using a WWW server and several CGI programs to facilitate access to the catalog database. This system provides the same accessibility to users from both inside and outside the university.

In recent months, the monthly access has reached 600 hits to the server.

2 Media Center

Following its establishment in the first quarter of this century, the OCU library gained nationwide recognition as a premier library facility. However, in recent years, it found itself unable to keep pace with the information revolution and thus to meet contemporary demand. The urgent need arose to modernize the facilities in order to promote educational and research activities at the university. Against this background, it was decided to integrate the library, computer facility and education in information processing in order to provide the students and researchers access to the modern tools of the information age. As a result, in October 1996, the Osaka City University Media Center was founded with the aim of providing a high technology academic and research environment that is focused on the present and coming information age.

In addition to librarians and administrative staff, the Media Center also has a dozen research staff from four broad areas of specialization, namely, Database and Multimedia Systems, Computing Systems, Network Systems, and Library and Information Science. In addition to research in their respective areas of specialization, they also conduct lectures and training sessions and support the management of the Center's information infrastructure.

2.1 Facilities by Zone

2.1.1 Library Zone

The library zone occupies eight of the thirteen floors of the Media Center. There are two floors for deposit, one for periodicals, one for reference, two for open-shelves and reading, one for open-shelves and reading for researchers only, and one for rare books.

Each floor is about 2,800 square meters in area. Holdings number 2 million books, 6,000 periodical titles and 15,000 multimedia titles. This zone is equipped with several network access terminals to facilitate the retrieval of library information and provide access to the electronic library services mentioned in a later section.

2.1.2 Information Processing Zone

The information processing zone consists of a parallel server, high-speed graphics and image processing servers and other computing facilities. This zone also functions as a nucleus of the campus network and provides global Internet access to students and researchers.

2.1.3 Multimedia Zone

The multimedia zone provides facilities for using the wide variety of multimedia products that form part of library collection. In addition, the training room for information processing allows students freely to use computer facilities for their academic work. There is also a language laboratory with multimedia equipments.

2.2 Electronic Library Services

The following electronic library services are accessible through the Internet.

2.2.1 OPAC

Like university libraries, we also provide an OPAC service. Recently access has reached about one hundred thousand hits per month.

2.2.2 Secondary Information Database

Library information is also available on CD-ROM and is accessible from the campus LAN.

2.2.3 Corpus Database

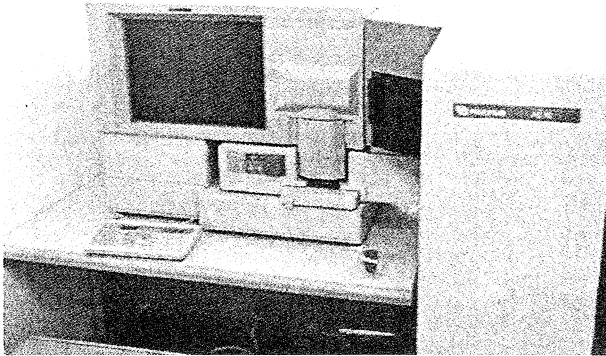


Figure 1. Exterior view of the MIRS

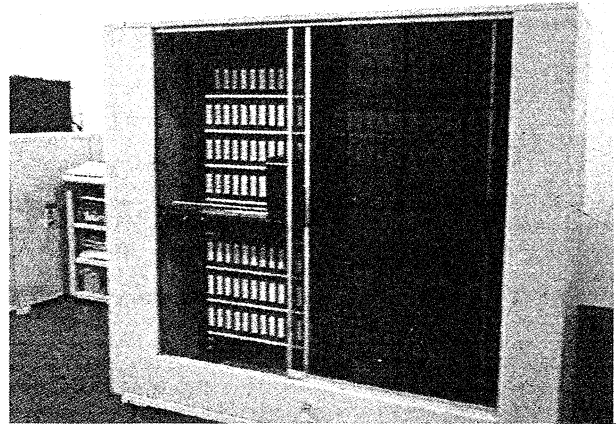


Figure 2. The automatic cartridge storage

We provide two kinds of corpus database. One is a database of the whole texts of journals published by institutes or faculties in the university. The other is a database of selected rare books.

2.2.4 MIRS

We also provide a service for image retrieval from microfilm collections. While the corpus database system requires all images to be stored in digital format in advance, the microfilm image information retrieval system (MIRS) uses images that are stored in analog format on microfilm. Further details are given in a later section.

2.2.5 Great Hanshin Earthquake Database

Almost three and half years ago, the great Hanshin-Awaji earthquake killed more than 6,000 people in the Kobe and Hanshin area and caused widespread and severe infrastructural damage. Osaka also suffered considerably. Osaka City University organized an academic research team that surveyed and recorded the damage from the viewpoint of geology. The Great Hanshin Earthquake Database consists of images recorded during the survey, aerial video images, and geographical information on Kobe. Images are linked to topographical maps on a web page.

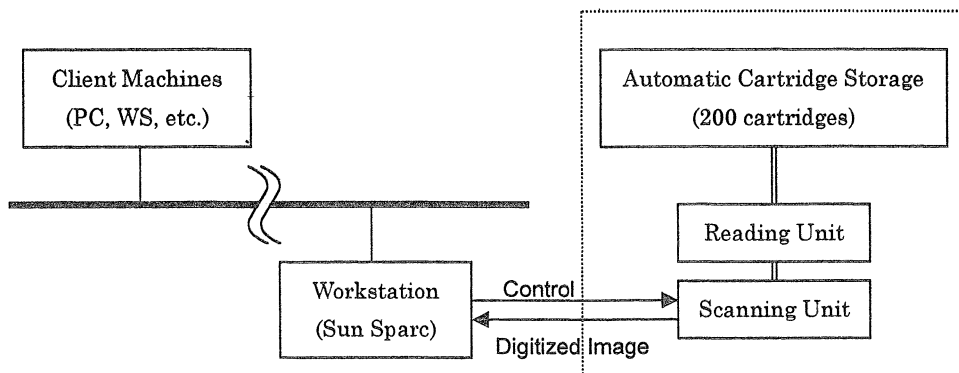


Figure 3. Hardware diagram of the MIRS

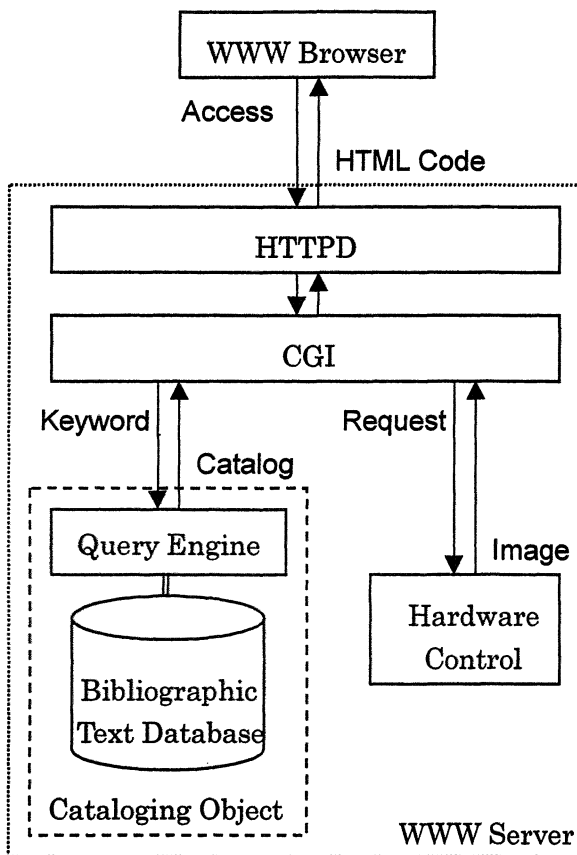


Figure 4. Software diagram of the MIRS

3 MIRS

3.1 Concept

The MIRS is a system of image retrieval from microfilms as mentioned briefly above. The main concepts of this system are as follows.

- (1) Combining catalog information and images:
Catalog information on images is input into a database. Keyword query to the database allows specification of the related frames on microfilms.
- (2) Long-term offering of microfilms:
The huge volume of microfilms already accumulated is immediately placed in service without prior conversion of the medium. Digital processing of an image

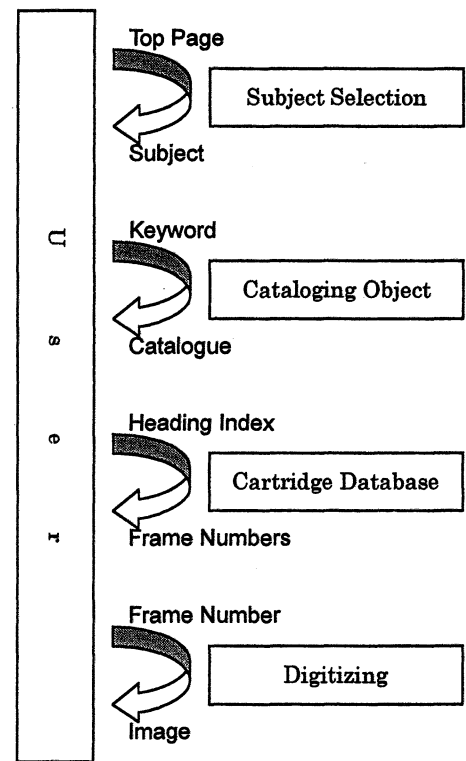


Figure 5. Retrieval protocol

upon request can provide various kinds of preliminary processing, such as enlargement, rotation, and enhancement.

- (3) Friendly user-interface:
Since users are neither specialists nor professionals in computer science, the user interface of the query system and the database construction must be easy to handle.
- (4) Flexible extension capability:
Other large-scale media like CD-ROM or Photo CD can be seamlessly incorporated into the system. Microfilm should be merely one of a variety of mass storage media.
- (5) Query functions independent of the platform:
Various functions of query and the structure of the database should be independent of the specific computer

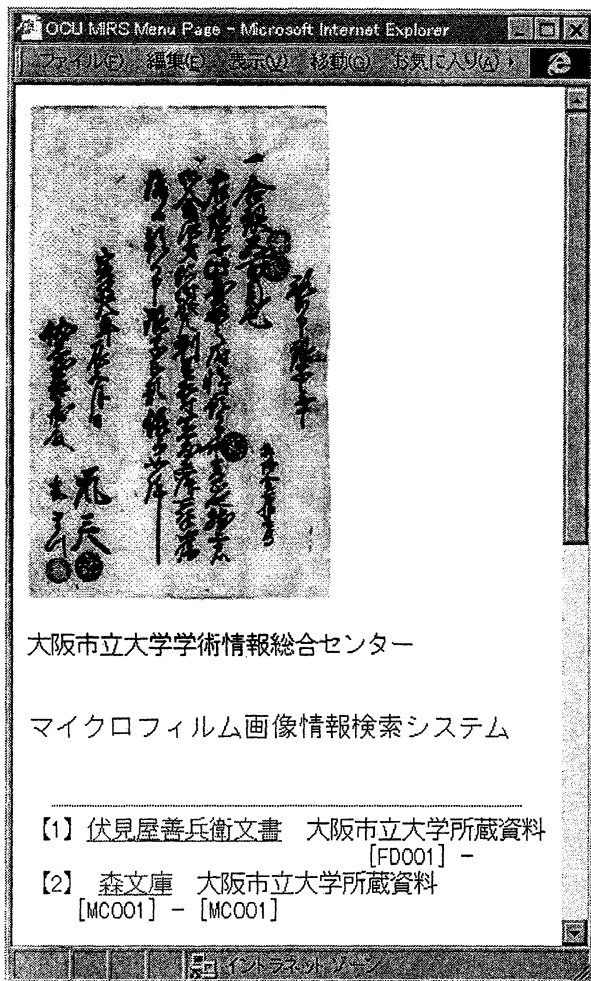


Figure 6. Menu page for selecting subject machine or operating system (OS).

3.2 Hardware

This system is characterized by the combination of a WWW server and a voluminous microfilm database. Its main components are (1) automatic cartridge storage for 16-mm microfilm; (2) an image printer which converts the optical analog image of microfilm into a digitized image; and (3) a WWW server workstation which controls the components above and distributes the image to the client (Figure 3). The maximum capacity of the MIRS is 200 cartridges or 1 million images.

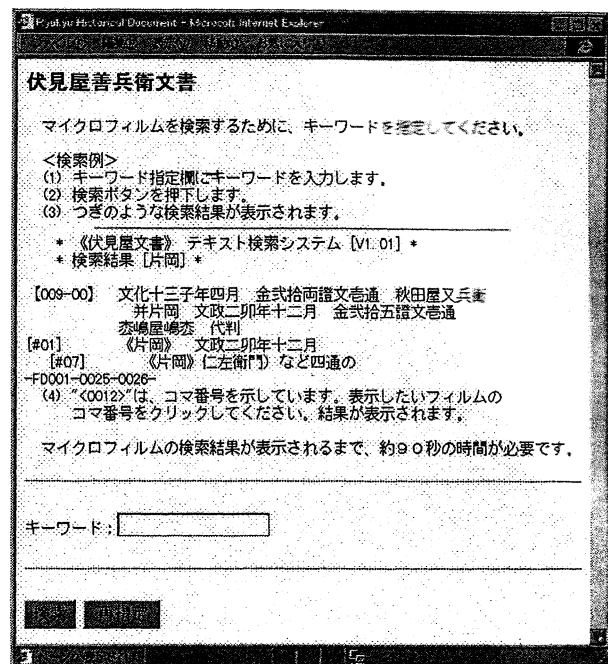


Figure 7. Page for keyword query

3.3 Software

To realize the concepts of the system, a WWW server is used. The client user can access the server through a WWW browser.

The software components of the server consist of CGI (Common Gateway Interface) programs. One component is for querying the cataloging object, and another is for controlling the digitizing hardware (Figure 4).

The retrieval protocol is shown in Figure 5. The user first selects the subject of the material (Figure 6). Then a page for keyword query is sent to the browser (Figure 7). After one or more keywords have been input, a catalog output relating to the keywords is transmitted (Figure 8). This catalog shows the headings, physical frame numbers underlined, and cartridge numbers. To proceed, the user should specify the frame number to retrieve. After a minute or so, the digitized image will be transmitted to the client through the Internet (Figure 9).

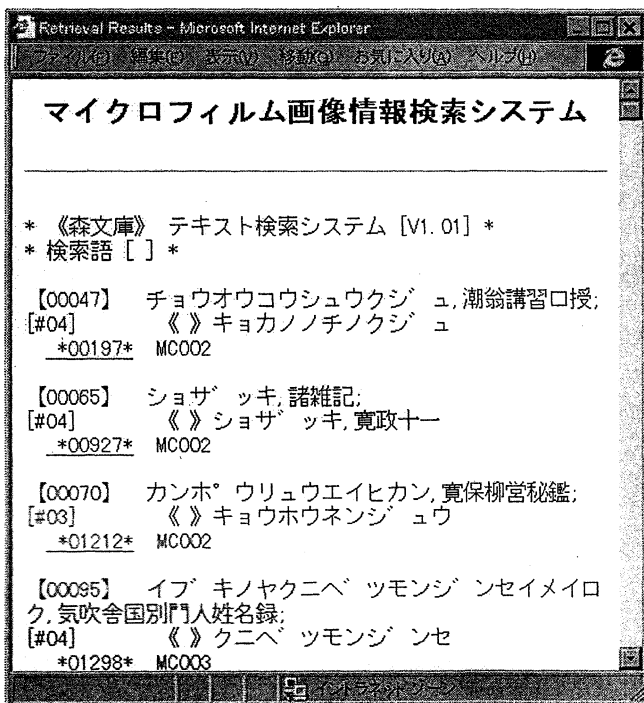


Figure 8. Catalog output

3.4 Services

Two collections are presently in service. One is a part of the Fushimiya Documents, a large quantity of manuscripts of economic transactions in the Edo period, from the 17th century to the 19th century. About 4,000 scenes selected from the Documents are stored in one microfilm cartridge. The other is the Mori Collection which includes biographies, history, literature, and topography of Japanese in the Edo period. About 300,000 scenes are stored in 94 cartridges. Table 1 shows the quarterly access records.

At present, the third collection, a collection of Ryukyu Material, is about to be imported.

4 Epilog

The MIRS, which digitizes an image on microfilm upon request, relies on mechanical equipment. This equipment and the microfilm constitute the potentially weak points of the system. However, the system has been working without serious trouble since it came into service almost two years ago.

We have been developing some advanced ideas to increase the reliability and accessibility. These are (1) automatic digitalization of consecutive scenes (2) CD-ROM as a mass storage medium and (3) a database query engine.

At present, the images retrieved from the MIRS are monochrome, but ideally images of historical materials should be colored. More technical innovations in the network will hopefully solve this point.

References

Shibayama, M., Namiki, M. 1996. An implementation of a retrieval system for the microfilm image databases through the World Wide Web (in Japanese). *IPSJ SIG Notes*, 96 (110), pp. 37 - 42.

Table 1. Quarterly access record of the MIRS

Period	Fushimiya Documents	Mori Collection
Oct - Dec 1996	2535 (868)	(not in service)
Jan - Mar 1997	503 (336)	878 (407)
Apr - Jun 1997	626 (304)	1050 (338)
Jul - Sep 1997	549 (403)	478 (267)
Oct - Dec 1997	482 (340)	709 (454)
Jan - Mar 1998	437 (340)	898 (497)
Apr - Jun 1998	429 (250)	619 (521)

(Unit: Page)

Note: Numbers in parentheses are accesses from outside of the university.

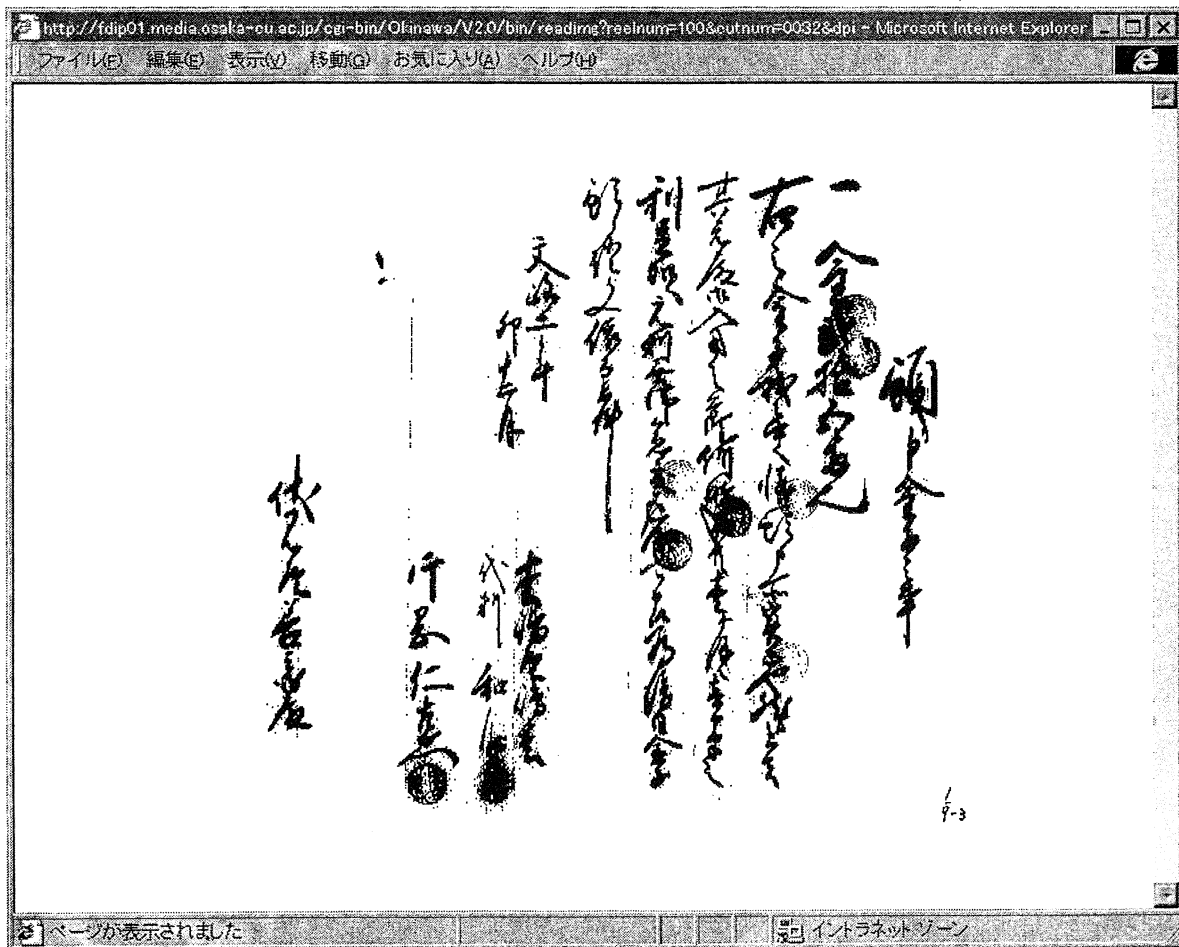


Figure 9. Retrieved image from the Fushimiya Documents

Shibayama, M. 1997. Kobunsho gazo detabesu [Image database of historical materials] (in Japanese). *Jinmongaku to Johoshori*, 15, pp. 45 – 50.

The Dublin Core Data Model

Eric Miller and Stuart Weibel
OCLC Office of Research, USA
Email :[emiller,weibel]@oclc.org

Abstract

The co-evolution of the semantic framework of the Dublin Core, together with the syntactic and structural facilities provided by XML and RDF, has reached the point where a new generation of Web-based resource description is possible. The additional functionality of these technologies over HTML-based metadata will support new services and opportunities that should enhance the state of resource description for electronic resources and improve prospects for cross discipline and international interoperability.

Cataloging for the Web: Dublin Core and the Resource Description Framework

Thomas Baker
Asian Institute of Technology
Thomas.Baker@cs.ait.ac.th

Abstract

Since the rise of the World Wide Web in 1994, libraries and information providers share an Internet Commons. As in a linguistically complex land, the content providers there use a wide variety of incompatible formats and cataloging rules to describe their information. But likewise since 1994, there has been a growing initiative on the part of national libraries, government agencies, publishers, and universities to develop a deliberately simple "pidgin" set of metadata (cataloging) elements that is interoperable with these various systems. This has resulted in the Dublin Core -- a set of fifteen basic categories with well-understood meanings like "Author" and "Title". This two-page standard has already been translated into nine languages. In parallel with the Dublin Core effort, the World-Wide Web community, together with software companies such as Netscape and Microsoft, has developed the Resource Description Framework (RDF) -- a standard format for encoding the metadata of Web documents, library catalogs, and electronic commerce applications. RDF uses XML, the simplified SGML that is positioned to replace HTML and word-processing formats as the standard format for documents on the Web. Together, the Dublin Core and RDF could provide a metadata system that is consistent across a wide range of applications and domains, no more complex than it needs to be, usable by both experts and non-experts, interoperable with existing library catalogs and legacy databases, and coherent across many languages. This tutorial will provide an introduction to the issues, tools, and prospects.

Introduction to DC-based Union Cataloging Systems for Academic Journals in Digital Library Environments

Han Suk Choi
Korea Research Information Center

Abstract

As rapidly growing digital library and Web-based information technology, it is not difficult for researchers to access online electronic collections over the world-wide publishers and information providers. However, the cost for online academic journals is so expensive and varying that most individual researchers can not afford to subscribe all of them. Korea Research Information Center(KRIC) established for providing the most up-to-date academic research information through union cataloging service, oversea online journal service, and specialized information service. KRIC has developed Research Information Service Systems(RISS) since 1997. One of the interesting RISS service is Dublin Core(DC)-based union cataloging systems for academic journals. We have designed and currently implemented the DC-based union cataloging systems for academic journals. In this presentation, we introduce the system architecture and design details of the DC-based union cataloging systems for academic journals. The major subsystems of DC-based union cataloging systems are server-side union catalogue management system, online shared cataloging service system using Z39.50 standard IR protocol, and copy-fax service system. The proposed DC-based union cataloging systems will share with most university libraries in Korea. Each local library generates bibliographic records of academic journals which they have, and then register the metadata of the academic journal records into the union catalog database in KRIC, which is geographically distributed from the local libraries. The major contents of this presentation are as follows;

1. Introduction
2. Metadata Elements for DC-based Union Cataloging Systems
3. DC-based Union Cataloging Systems for Academic Journals
 - System Architecture
 - DC-based Union Cataloging Management System
 - Online Shared Cataloging Service System
 - Copy-Fax Service System
4. Discussion
5. Conclusion

Extension of MHTML to Text Input and Text Search Functions in Multiple Languages on Off-the-shelf Browsers

Shigeo Sugimoto, Shigetaka Nakao, Myriam Dartois, Jun Ohta,
Akira Maeda*, Tetsuo Sakaguchi, Koichi Tabata

University of Library and Information Science
Tsukuba, Ibaraki, Japan
{sugimoto, nakao, myriam, jun, saka, tabata}@ulis.ac.jp

*Nara Institute of Science and Technology
Ikoma, Nara, Japan
aki-mae@is.aist-nara.ac.jp

Abstract

The World Wide Web (WWW) covers all over the world. However, browsing function for documents in multiple languages is not widely available yet for casual users. From the viewpoint of digital libraries, functions to display and input multilingual texts are obviously crucial. Multilingual HTML (MHTML) is a browser technology for multilingual documents on WWW. The authors developed a display function of multilingual documents based on the MHTML technology. This technology was intended as a simple, light, easy-to-use and inexpensive technology to view multilingual documents via the Internet. The authors have extended it to text input function in multiple languages on off-the-shelf browsers. The new technology gained by this extension is quite useful to create an environment for end-users of digital libraries where they are able to view and search multilingual documents

from any off-the-shelf browser. In addition to the extended MHTML technology, this paper shows an SGML-based full-text retrieval system which is developed using the extended MHTML. It has a user interface to input queries and to display results in multiple languages.

Keywords

Multilingual Document Browsing, Off-the-Shelf WWW Browsers, Multilingual Texts Display and Input, Text Retrieval in Multiple Scripts

1 Introduction

The Internet and the World Wide Web (WWW) are very important infrastructure for digital libraries. English is widely accepted as a common language on WWW for global communication, but on the other

hand, there are a huge amount of documents written in non-English languages on WWW. It is obvious that functions to access information in foreign languages as well as English are crucial for WWW and digital libraries. From another viewpoint, since libraries are inherently multilingual, library information systems have to cope with multilingual library information. In the case of library information systems handling Chinese, Japanese and Korean (CJK) texts, display and input functions for a large set of characters containing non-standardized characters is one of the key technologies to build a digital library.

The authors have been working on a browser technology named Multilingual HTML (MHTML) to display multilingual documents on an off-the-shelf WWW browser even if the browser has no fonts required to display the documents[4][5][7]. Since a document browsing function is realized as a Java applet, users are required to have only an off-the-shelf browser which is capable to run Java applets, i.e., a browser applet running on a browser. We have applied the MHTML technology to a gateway service to view foreign documents and to a multilingual electronic text collection of folktales[2][3]. We have extended MHTML to realize a text input function in multiple languages. We have implemented a Japanese text input server which sends a user interface applet to input Japanese words/characters from a remote client without any Japanese functions. Since the extended MHTML technology is designed independently of languages, it is extensible to other languages. We have also applied the extended MHTML to an SGML-based text retrieval system, which is a full-text database for documents written in multiple languages and has a user interface built based on the MHTML technology.

2 Display and Input of Texts in Multiple Languages

A display function of HTML documents and a text input function on a client are the most basic functions required to access information on WWW. However, these functions for texts in foreign languages are not always provided on a client. The MHTML project had initially started to realize a light, easy-to-use and ubiquitous environment to browse WWW documents written in multiple languages on an off-the-shelf WWW browser. We developed a viewing function to display multilingual documents on a client where fonts for multilingual texts are not necessarily installed. Key aspect of MHTML was that, even if a standard character code set for multilingual texts is widely accepted it is not practical to assume that end users of digital libraries can afford a complete set of fonts for the code set.

Text input function is crucial as well as the display function for browsing documents in multiple languages. The text input function is generally defined as a mapping to a character code or a code string from a user action on an input device, i.e., a single keystroke, a combination of keystrokes, a sequence of keystrokes, a mouse click, and so on. Since inputted texts are usually displayed on a screen, the text input function requires a display function as well. In the case of an ordinary Japanese text input function, for example, a Japanese word or phrase expressed in phonetic characters (Hiragana, Katakana or alphabets) is converted to an appropriate Japanese word or phrase expressed in Kanji, Hiragana, Katakana and/or alphabets. The character code string emitted from the function is displayed on a screen using font locally installed. The phonetic expression in al-

phabets, i.e., transliteration, can be used to input texts from a conventional ASCII keyboard. The mapping function can be located in the client or in a server connected via a network, but the font has to be locally provided. In addition, users have to set up their local environments in accordance with the requirement to input texts such as connection to the mapping function and font installation. It is difficult for an end user who casually accesses foreign documents to set up his/her environment.

3 MHTML

3.1 Basic Concepts

An MHTML server and an MHTML object are the key components of the MHTML technology. The MHTML server fetches a document from a document server, converts it into an MHTML object and sends the object to a client with an applet to display the object on the client. The MHTML object contains the text string of the source document, and a minimum set of font glyphs required to display the text. (The character string in the object is internalized object by object, so that it can not be re-converted to the source character code string.) Since the applet can display all of the characters contained in the source document using only the glyphs sent from the server, the client need no font for foreign languages. Figure 1 shows an MHTML object. The repertoire of the languages of the server primarily depends on the set of fonts stored in the font bank because conversion of documents into ISO-2022-JP-2[6] standard is usually straightforward.

MHTML also has the advantage that it can display a document which contains a non-standard characters. These non-standard characters occasionally appear in Japanese texts and they are called Gaiji

in Japanese. A Gaiji can be assigned a code and a glyph locally, but the code has no meaning outside the local machine. An MHTML server can display any character if it is given its glyph and code. There are two ways to make a Gaiji displayable on a client, (1) to add new glyphs to a font file in the font bank, and (2) to add a new font file, which contains glyphs for a set of Gaijis, to the font bank.

3.2 Extension of MHTML – Text Input

Text search is a primary function for information access, i.e., text search in a database and in a document displayed on a screen. Text input function in multiple languages is indispensable to realize text search function in multiple languages. By slightly extending the MHTML technology, we have gained a framework for text input in multiple languages from an off-the-shelf browser. As illustrated in Figure 2, the extended MHTML object contains an identifier of character encoding and character codes in addition to the components of a basic MHTML object shown in Figure 1. A source character code string can be reproduced from an extended MHTML object by replacing every character in the internalized text string by its corresponding source character code. The character encoding identifier is required to make the reconverted text conformant with the ISO-2022-JP-2 standard. Conversion to Unicode[8] is also possible. Based on the framework, we have developed a Japanese text input function for an off-the-shelf browser running on a client which has no Japanese text environment.

Figure 3 shows an outline of a Japanese text input server based on MHTML. The Japanese input server has been implemented using a text input software called Wnn and a user interface applet defined

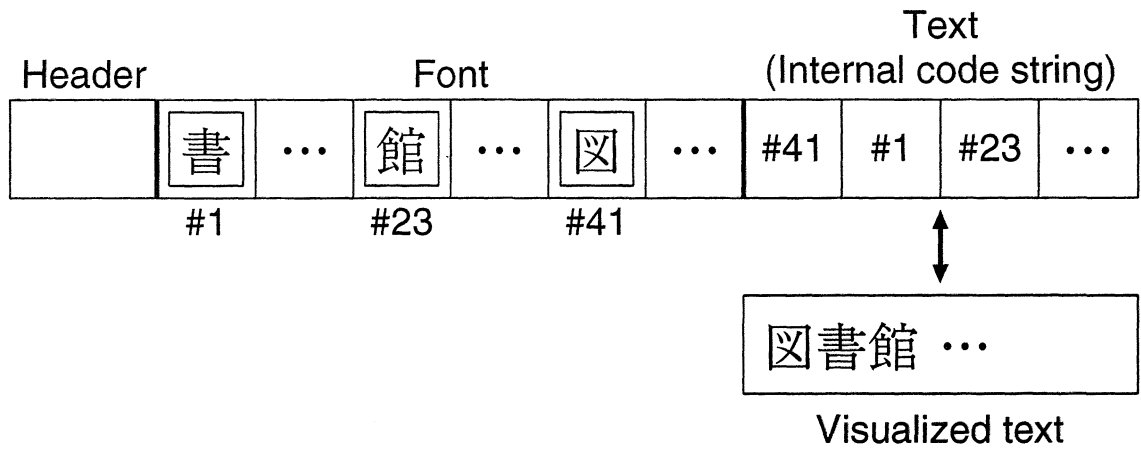


Figure 1: MHTML object

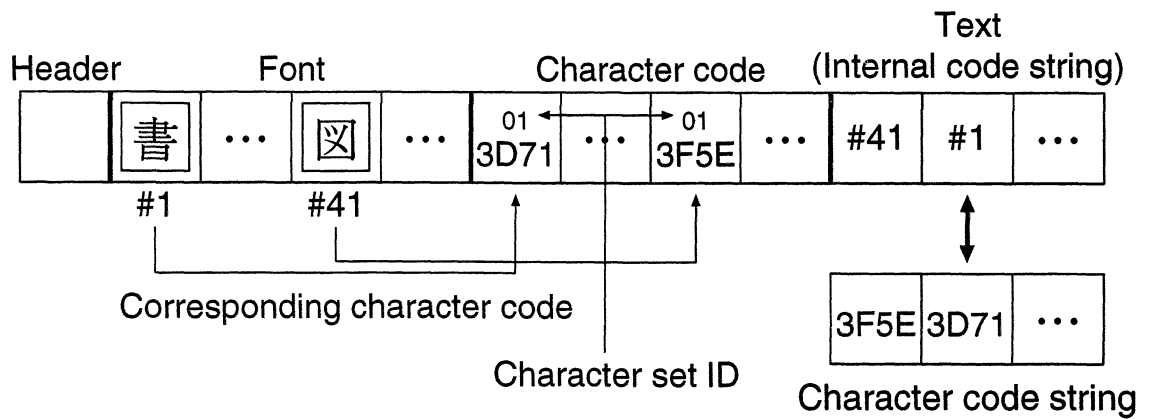


Figure 2: Extended MHTML object

based on the extended MHTML. The text input server (TI server) is located between a client and a WWW server. The TI server receives a Japanese word written in a transliterated form and produces a list of Japanese words. Figure 4 shows a user interface applet to input Japanese texts. This text input applet has a text input field to type a Japanese word or phrase in transliterated form. The CONVERT button means to send the inputted string to the Japanese text input server. A list of words is returned from the TI server and is displayed on the left of the input field. A word selected by a mouse click on the list is displayed in the top line. Users can append additional words/characters to form a complete word or phrase for retrieval. The SUBMIT button means to send the character code string displayed in the top line to a text search function which uses the string for retrieval. The Japanese texts displayed on the applet are sent as an extended MHTML object.

4 An SGML-based Text Retrieval System

We have experimentally developed an SGML-based text retrieval system which is designed to receive texts encoded in multiple scripts. Currently, it can store and retrieve Japanese and ASCII texts, but it is extensible to texts scripted in any character codes. Its user interface is designed using MHTML in order to provide users with ubiquitous accessibility to texts scripted in multiple languages. Figure 5 shows the user interface of the text retrieval system. The window at the bottom is a text input window shown in figure 4. The window at the top shows a list of hits gained by a retrieval. A user can display a source text by clicking on an element of the list.

Figure 6 shows the outline of the system

configuration. It receives a text encoded in a regional standard and converts it into the ISO-2022-JP-2 standard. The text is converted into a Unicode-based text to create the index. And, on one hand, the text is stored as it is. ISO-2022-JP-2 standard has multiple character code spaces for character sets of regional languages and defines switching protocol between the spaces. This feature is quite advantageous to use an existing document encoded in a regional standard in the internationalized environment. However, this encoding scheme is disadvantageous to make a text retrieval function simple. On the other hand, since the flat character space given in Unicode is advantageous for a simple text retrieval function, Unicode is employed as the basic encoding scheme to build the index. The index is created based on N-gram. Its user interface is created using the extended MHTML. Thus, this system provides a framework for retrieving texts encoded in multiple encoding scheme (i.e., multilingual documents) with ubiquitous user interface for retrieval.

5 Conclusion

The technology implemented as MHTML is quite simple. The research of MHTML was started to realize a simple, light, easy-to-use and inexpensive environment to read and write foreign texts in the WWW environment. The functions implemented in the research have proved the feasibility of such environment. We believe that the framework realized in MHTML has potential to change paradigm of text input and output in a distributed environment.

The authors have applied the MHTML technology to build user interfaces for an electronic text collection, an OPAC and a full-text retrieval system. They are also collaborating with the internationalization

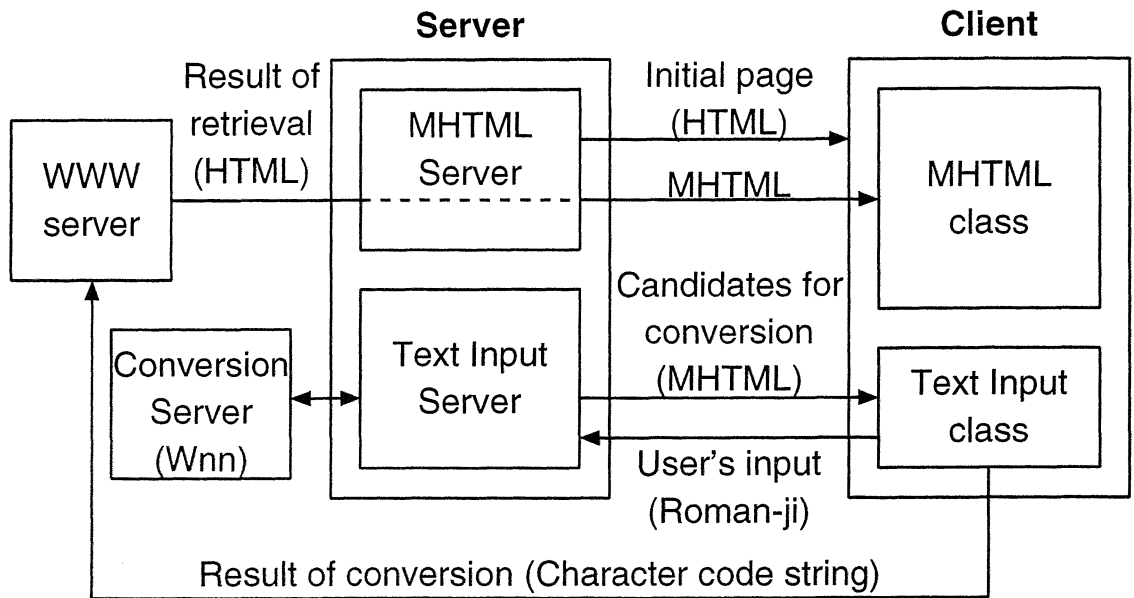


Figure 3: Text Input Server

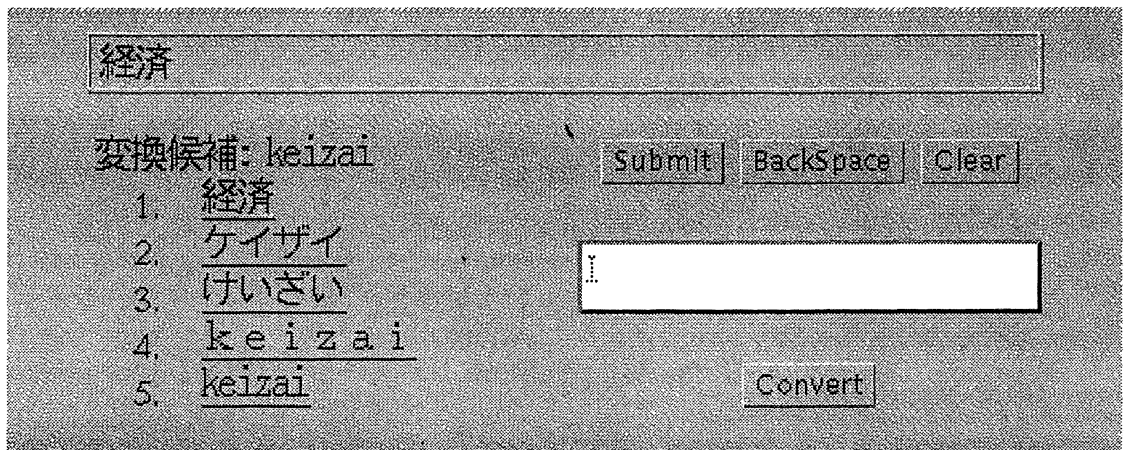


Figure 4: Text Input Applet

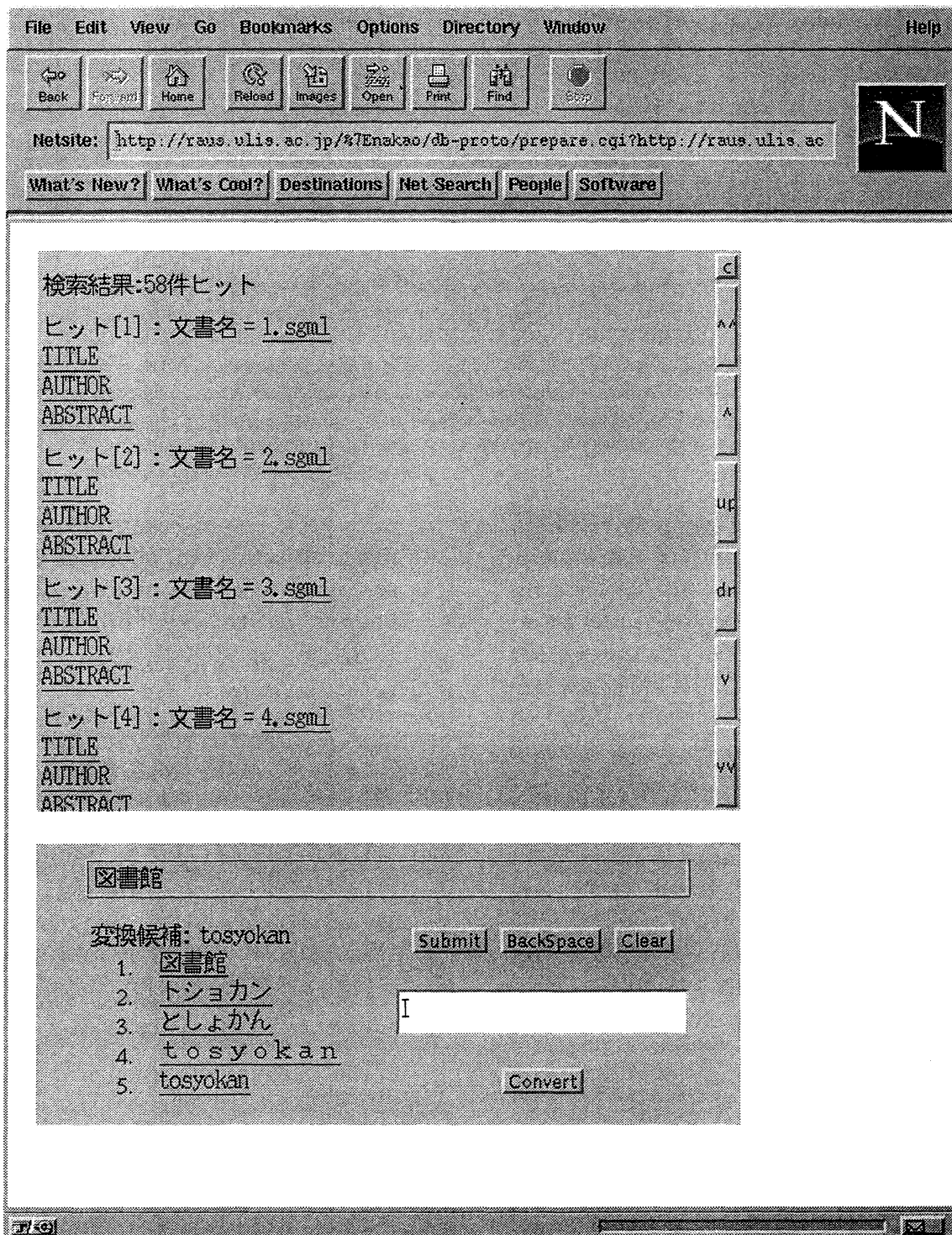


Figure 5: User Interface of the SGML-based Text Retrieval System

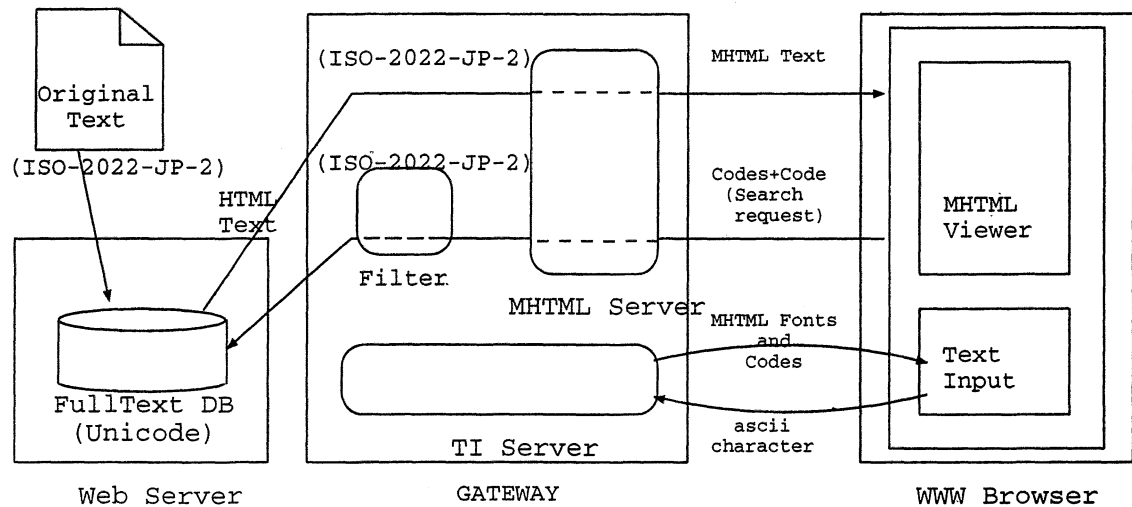


Figure 6: Overview of an SGML-based Full-text Retrieval System for Multilingual Documents

group of Dublin Core[1] to apply MHTML for multilingual user interface for a meta-data database. These application systems require rather simple text-based user interfaces but not fancy ones. To extend the repertoire of languages, MHTML requires fonts. Fonts of public domain are required to extend services for not-for-profit services.

References

- [1] Baker, T. and Weibel, S.; Dublin Core in Thai and Japanese: Managing Universal Metadata Semantics, Digital Libraries, no.11, pp.35-47, 1998 (in Japanese)
- [2] Dartois, M., et al.; A multilingual electronic collection of folk tales for casual users using off-the-shelf browsers, D-lib magazine, 1997, <http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html>
- [3] Dartois, M., et al.; Building a multilingual electronic text collection of folk tales as a set of encapsulated document objects: An approach for casual users to browse multi-lingual documents on the fly, Proceedings of ECDL'97, pp.215-231, 1997
- [4] Maeda, A., et al.; Viewing Multilingual Documents on Your Local Web Browser, CACM, vol.41, no.4, pp.64-65, 1998
- [5] Maeda, A., et al.; A Multilingual HTML Document Browsing System for Clients without Multilingual Fonts, Transactions of IPSJ, vol.39, no.3, pp.802-809, 1998 (in Japanese)
- [6] Ohta, M. and Honda, K.; ISO-2022-JP-2; Multilingual Extension of ISO-2022-JP, RFC 1554, 1993
- [7] Sakaguchi, T., et al.; A Browsing Toll for Multi-lingual Documents for Users without Multi-lingual Fonts, Proceedings of DL'96, pp.63-71, 1996
- [8] The Unicode Consortium; The Unicode Standard, Ver.2.0, Addison-Wesley, 1996

Reasoning about RDF Elements

Vilas Wuwongse¹, Chutiporn Anutariya¹ and Ekawit Nantajeewarawat²
Email: vw@cs.ait.ac.th, ca@cs.ait.ac.th and ekawit@siit.tu.ac.th

¹Computer Science & Information Management Program
School of Advanced Technologies
Asian Institute of Technology
Pathumtani 12120, Thailand

²Department of Information Technology
Sirindhorn International Institute of Technology
Thammasat University
Pathumtani 12120, Thailand

Abstract

This paper proposes a theoretical framework for reasoning about RDF elements by employment of the theory of Declarative Programs. With reasoning capability, rules can be defined in order to explicitly and formally state semantic relationships between RDF elements, which will enable inference or derivation of new RDF elements from existing ones. Hence, in addition to computation by pattern matching, the proposed framework can formulate and handle complicated computations with RDF elements. Application of this framework to Web-based resource discovery problems is presented with an example.

Keywords: RDF elements, declarative programs, resource discovery problems.

1 Introduction

The recent expansion in the use of the Internet and the World Wide Web as information distribution media has made available huge resources in electronic form, so that today's Web resembles a gigantic library of books, pictures, songs, movies, weather forecasts, horoscopes, yellow pages, etc. The development of an effective information retrieval technique has become one of the major issues in view of the fact that almost all presently available search engines merely employ techniques which simply match words or sentences in documents. Users are often annoyed at receiving several thousands hits after just typing in a few keywords. Moreover, these search results still do not present users with enough information to determine their relevance, so that, in turn, in most cases further exploration is required.

A solution, proposed to date for coping with these limitations on the Web, provides sufficient descriptions, known as *metadata*. A search by means of metadata is expected to be more efficient and more accurate, since users can precisely formulate more specific queries which will yield more precise answers. For example, a user may want to find any items authored by John Smith, i.e., only items authored by John Smith rather than all items containing the words John Smith.

Relying on the concepts of metadata, the *Resource Description Framework (RDF)* [5], a collaborative design effort under the auspices of the W3C (World Wide Web Consortium), has been developed, in order to provide a unified framework for processing metadata. RDF metadata can be adopted in a wide range of application areas including resource discovery, digital library and electronic commerce.

This paper attempts to develop a framework and technique for reasoning about RDF elements. With this capability, one can define rules, which describe semantic relationships between RDF elements, and will be able to derive new RDF elements from existing ones even though they have not previously explicitly stated. In addition, application of the proposed framework to resource discovery problems allows to incorporate reasoning or deductive capabilities into the retrieval process.

The proposed framework employs *Declarative Program (DP) Theory* [1, 2, 3] - a generalization of the conventional logic program theory, in which terms are generalized into any *data objects* and substitutions into *specializations*. In particular, it has been carefully defined with generality and applicability to data structures of any domains, each of which is characterized by a mathematical structure, called a *specialization system*.

Section 2 introduces the DP theory, Section 3 develops a specialization system for RDF elements and RDF Declarative Programs, Section 4 presents an application of RDF Declarative Programs to Web-based Resource Discovery Problems with an example and Section 5 draws conclusion.

2 Declarative Program Theory

Certain fundamental definitions of declarative program theory [3] will be recalled.

2.1. Specialization Systems

A *specialization system* is an abstract structure derived from the generalization of *substitutions* in the conventional logic programs, and defined in terms of certain very simple axioms:

Definition 1 (Specialization System)

A *specialization system* is a four-tuple $\Gamma = \langle \mathcal{A}, \mathcal{G}, \mathcal{S}, \mu \rangle$ of three sets $\mathcal{A}, \mathcal{G}, \mathcal{S}$ and a mapping μ from \mathcal{S} to the set of all partial mappings on \mathcal{A} , i.e., $\mu: \mathcal{S} \rightarrow \text{partial_map}(\mathcal{A})$, that satisfies the requirements:

1. $\forall s_1, s_2 \in \mathcal{S} \exists s \in \mathcal{S}: \mu(s) = \mu(s_1) \circ \mu(s_2)$,
2. $\exists s \in \mathcal{S} \forall a \in \mathcal{A}: \mu(s)(a) = a$,
3. $\mathcal{G} \subset \mathcal{A}$

where $\mu(s_1) \circ \mu(s_2)$ is the composite mapping of the partial mappings $\mu(s_1)$ and $\mu(s_2)$. The elements of \mathcal{A}, \mathcal{G} and \mathcal{S} are called *objects*, *ground objects*, and *specializations*, respectively, the set \mathcal{G} the domain, and the specializations, which satisfy the second requirement, *identity specializations*. \square

Intuitively, Conditions 1 to 3 mean that

1. for all specializations s_1 and s_2 , there exists a specialization s such that the corresponding partial mapping of s is the composition of the two mappings corresponding to s_1 and s_2 ,
2. there is a specialization that does not change any objects, and
3. ground objects are objects.

For $\theta \in \mathcal{S}$, postfix notation allows to represent $\mu(\theta)(a)$ as $a\theta$. If b exists such that $\mu(\theta)(a) = b$, θ is said to be applicable to a , and a specialized to b by θ . A specialization $\theta \in \mathcal{S}$ is applicable to a subset B of \mathcal{A} if θ is applicable to any element of B ; $B\theta$ is then defined by $B\theta = \{b\theta \mid b \in B\}$. Henceforth, any condition that includes $a\theta$ or $B\theta$ will also include the condition that θ is applicable to a or B .

Definition 2 (Mapping yielding Sets of all Ground Objects)

A mapping $\text{rep}: \mathcal{A} \rightarrow \text{powerset}(\mathcal{G})$ is defined by $\text{rep}(a) = \{g \mid g = a\theta \in \mathcal{G}, \theta \in \mathcal{S}\}$. \square

2.2. Declarative Programs

Declarative programs and other related concepts can now be defined in terms of a specialization system $\Gamma = \langle \mathcal{A}, \mathcal{G}, \mathcal{S}, \mu \rangle$.

Definition 3 (Declarative Program)

Let X be a subset of \mathcal{A} . A *definite clause* on X is a formula of the form:

$$H \leftarrow B_1, B_2, \dots, B_n$$

where H, B_1, B_2, \dots, B_n are *objects* in X , often referred to as *atoms*. H is called the *head* and (B_1, B_2, \dots, B_n) the *body* of the *definite clause*. The set of all *definite clauses* on X is denoted by $D\text{clause}(X)$. A *declarative program* on X is a (possibly infinite) set of *definite clauses* on X . A *declarative program* on \mathcal{A} is also called a *declarative program* on Γ . \square

A definite clause will often be called simply a clause. Let C be a definite clause. The head of C will be denoted by $\text{head}(C)$, and the set of all atoms in the body of C by $\text{body}(C)$. A clause C is called a *unit clause* if $\text{body}(C)$ is empty.

Let C be a definite clause $C = (H \leftarrow B_1, B_2, \dots, B_n)$. When $\theta \in \mathcal{S}$ is applicable to H, B_1, B_2, \dots, B_n , a definite clause $C\theta = (H\theta \leftarrow B_1\theta, B_2\theta, \dots, B_n\theta)$ is obtained. A definite clause C' is an instance of C iff there is a specialization θ such that $C' = C\theta$. A definite clause C is a ground definite clause iff $C \in D\text{clause}(\mathcal{G})$. A ground instance of a clause C is a ground clause which is an instance of C . Let P be a declarative program on

Γ . Denote the set of all ground instances of definite clauses in P by $Gclause(P)$.

2.3. Semantics of Declarative Programs

The mapping $T_P: 2^{\mathcal{G}} \rightarrow 2^{\mathcal{G}}$, defined for a program P on Γ , will be used to define the declarative semantics of a program in Definition 5.

Definition 4 (Mapping T_P)

Let $\Gamma = \langle \mathcal{A}, \mathcal{G}, \mathcal{S}, \mu \rangle$ and $I \subset \mathcal{G}$. A mapping $T_P: 2^{\mathcal{G}} \rightarrow 2^{\mathcal{G}}$ is defined by

$$T_P(I) = \{head(C) \mid C \in Gclause(P), body(C) \subset I\}.$$

By means of the mapping T_P , define declarative semantics of a program P :

Definition 5 (Declarative Semantics of a Program)

Let P be a program on Γ . The declarative semantics of P , $\mathcal{M}(P)$, is defined by

$$\mathcal{M}(P) = \bigcup_{n=1}^{\infty} [T_P]^n(\emptyset), \text{ where } \emptyset \text{ is the empty set.}$$

3 RDF Declarative Programs

3.1. Basic Patterns of RDF Elements

In the RDF Model and Syntax Specification proposed by the W3C, RDF elements are defined either in Serialization or Abbreviated Syntax, both of which are based on XML [4]. Serialization Syntax can express the full capabilities of the RDF data model, while Abbreviated Syntax includes additional grammar, in order to provide a more compact form for representation of a subset of the data model. Only Serialization Syntax will be considered in this paper.

It is assumed here that the namespace name for the RDF schema has been declared and abbreviated as RDF. Examples of RDF tag names are RDF:RDF, RDF:Description, RDF:Seq, RDF:Bag and RDF:Alt.

RDF elements have the form $\langle \text{RDF:RDF} \rangle \text{content} \langle / \text{RDF:RDF} \rangle$, where content is an

RDF term or a sequence of RDF terms. By convention, an RDF term assumes one of the forms:

(a) *empty form*

$$\langle t \ b_1="c_1" \ \dots \ b_n="c_n" \ / \rangle$$

(b) *simple form*

$$\langle t \ b_1="c_1" \ \dots \ b_n="c_n" \rangle \ c_{n+1} \ \langle / t \rangle$$

(c) *nested form*

$$\langle t \ b_1="c_1" \ \dots \ b_n="c_n" \rangle a_1 \ a_2 \ \dots \ a_m \ \langle / t \rangle,$$

where t is a tag type, the b_i are all distinct attribute names, the c_i are constants and the a_i are ground RDF terms. In order to represent implicit information contained in an RDF term, an RDF term used here may carry variables. The formal definition of RDF terms (with variables) will be given in the next sub-section.

Note that RDF terms of the empty form $\langle t \ b_1="c_1" \ \dots \ b_n="c_n" \ / \rangle$ can be equivalently represented in the simple form $\langle t \ b_1="c_1" \ \dots \ b_n="c_n" \rangle \ \langle / t \rangle$. In the sequel, all empty terms will be written in the simple form.

3.2. Specialization System for RDF Terms

This sub-section defines a specialization system for RDF terms, $\Gamma_R = \langle \mathcal{A}_R, \mathcal{G}_R, \mathcal{S}_R, \mu_R \rangle$, which will be used in the definition of the specialization system for RDF elements in Sub-section 3.3.

In the sequel, let T be a set of *tag types*, B *attribute names* (or *qualifiers*), C *constants*, $TVAR$ *tag-variables* (or *T-variables*), $BVAR$ *attribute-variables* (or *B-variables*), $CVAR$ *constant-variables* (or *C-variables*), and $PVAR$ *attribute-value-pair-variables* (or *P-variables*). Assume that:

1. B contains an element CONTENT.
2. C contains ϵ , which denotes the empty string or white spaces.
3. No element in T begins with "TVAR:" and every element in $TVAR$ begins with "TVAR:".
4. No element in B begins with "BVAR:" and every element in $BVAR$ begins with "BVAR:".
5. No element in C begins with "CVAR:" and every element in $CVAR$ begins with "CVAR:".
6. Every element in $PVAR$ begins with "PVAR:".

Definition 6 (RDF term)

An *RDF term* is a formula of the form:

$$\langle t P_1 \dots P_n \rangle \langle /t \rangle,$$

where

- t is a tag type in T or a tag variable in $TVAR$,
- $n \geq 0$ and P_i is a P-variable in $PVAR$ or P_i takes the form $b=c$, where
 - $b \in BVAR$ and $c \in C \cup CVAR \cup \mathcal{A}_R$, or
 - $b \in B - \{\text{CONTENT}\}$ and $c \in C \cup CVAR$, or
 - $b = \text{CONTENT}$ and $c \in C \cup CVAR \cup \mathcal{A}_R$.

The order of the P_i , called *attributes* is immaterial. Duplicate attributes are not differentiated¹. Moreover, attribute names other than **CONTENT** can not appear more than once in the same tag. \square

Note that:

1. Terms taking the form of:

$$\langle t P_1 \dots P_{i-1} \text{CONTENT}="c" P_{i+1} \dots P_n \rangle \langle /t \rangle,$$

where $c \in C$, and

$$\langle t P_1 \dots P_{i-1} \text{CONTENT}=a P_{i+1} \dots P_n \rangle \langle /t \rangle,$$

where $a \in \mathcal{A}_R$,

are normally written as:

$$\langle t P_1 \dots P_{i-1} P_{i+1} \dots P_n \rangle c \langle /t \rangle,$$
 and

$$\langle t P_1 \dots P_{i-1} P_{i+1} \dots P_n \rangle a \langle /t \rangle,$$

respectively. RDF terms taking the first form are called *simple terms* and those in the second form are *nested terms*. In addition, simple terms with no attributes or those taking the form

$$\langle t \rangle \langle /t \rangle$$

are said to be in *atomic form* and called *atomic terms*.

2. The distinctions between T-variables, B-variables, C-variables and P-variables are:

- T-variables are those that start with the prefix "TVAR:" and can only be specialized to RDF terms in \mathcal{A}_R ,
- B-variables are those that start with the prefix "BVAR:" and can only be specialized to attribute names in B ,
- C-variables are those that start with the prefix "CVAR:" and can only be specialized to constants in C , and
- P-variables are those that start with the prefix "PVAR:" and can only be specialized to sets of P-variables in 2^{PVAR} or pairs of attribute and value.

Definition 7 (\mathcal{A}_R , the set of RDF terms)

\mathcal{A}_R is the set of all *RDF terms* constructed from the elements in $T, B, C, TVAR, BVAR$ and $PVAR$. \square

Definition 8 (RDF tree)

An *RDF tree* is an equivalent graphical representation of an RDF term

$$\langle t P_1 \dots P_n \rangle \langle /t \rangle,$$

where t is shown as an ellipse node and the P_i as immediate subtrees of the node t denoted by *sub_i* (see Figure 1-a); if

1. P_i is a P-variable, the subtree representing P_i will consist of a single circle node representing P_i (see Figure 1-b),
2. P_i takes the form $b=c$,
 - 2.1. if c is a constant or a C-variable, the subtree representing P_i has as its root a circle node representing b , connected to a child rectangle node representing c (see Figure 1-c),
 - 2.2. if c is an RDF term, the subtree representing P_i has as its root a circle node representing b , connected to an RDF tree representing the term c (see Figure 1-d).

Denote the RDF tree of the term a by *tree(a)*. \square

By the definition of RDF trees, a particular RDF tree, representing the term a , is considered to be a 3-leveled tree, which contains

- at the first level, an ellipse node, representing the tag type or T-variable of the term a ,
- at the second level, circle nodes representing attribute names, B-variables and P-variables of a ,

¹ Note that two attributes $b = v_1$ and $b = v_2$, where $v_1 \neq v_2$ are not considered to be duplicate.

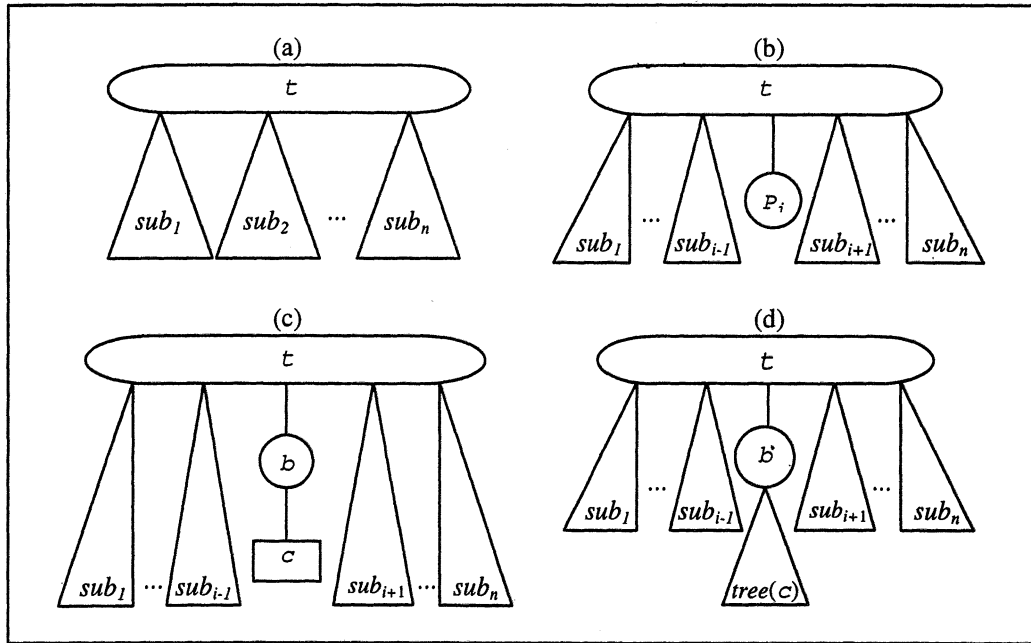


Figure 1. The RDF tree.

- at the third level, rectangle nodes and immediate subtrees, which, respectively, represent atomic values, C-variables and other terms nested in a ; in particular, these rectangle nodes and subtrees provide corresponding values for the attribute names and B-variables at the second level.

Definition 9 (\mathcal{G}_R , the set of ground RDF terms)

\mathcal{G}_R is that subset of \mathcal{A}_R which consists of all ground (variable-free) terms in \mathcal{A}_R . \square

In other words, \mathcal{G}_R is the set of ground RDF terms, assuming one of the forms:

$\langle t \ b_1="c_1" \dots b_n="c_n" \ \text{CONTENT}="c_{n+1}" \rangle \langle /t \rangle$
or
 $\langle t \ b_1="c_1" \dots b_n="c_n" \ \text{CONTENT}=\mathcal{g}_1 \dots \text{CONTENT}=\mathcal{g}_m \rangle \langle /t \rangle$,

where

- t is a tag type in T ,
- the b_i are all distinct attribute names in $B - \{\text{CONTENT}\}$,
- the c_i are constants in C , and
- the \mathcal{g}_i are RDF terms in \mathcal{G}_R .

The following notion of an expandable term will be used in the definition of the specialization mapping, ν_R (Definition 11).

Definition 10 (Expandable term)

Let the EXP attribute be the attribute-value pair written as EXP="here". Define an *expandable term* as a term t in \mathcal{A}_R which contains the EXP attribute. Then, let \mathcal{A}_R be the subset of \mathcal{A}_R that consists of all expandable terms in \mathcal{A}_R . \square

Definition 11 (ν_R , a specialization mapping)

Let \mathcal{E}_R be $(BVAR \times (B - \{\text{CONTENT}\})) \cup (BVAR \times \{\text{CONTENT}\}) \cup (CVAR \times C) \cup (PVAR \times (BVAR \times CVAR)) \cup (PVAR \times (\{\text{CONTENT}\} \times \mathcal{A}_R)) \cup (PVAR \times 2^{PVAR}) \cup (TVAR \times T) \cup (TVAR \times \mathcal{A}_R) \cup (TVAR \times \{\varepsilon\})$, where 2^{PVAR} is the power set of $PVAR$. The specialization mapping $\nu_R: \mathcal{E}_R \rightarrow \text{partial_map}(\mathcal{A}_R)$ is defined as follows:

1. *Attribute Name Restriction*

- When $e = (bvar, b) \in BVAR \times (B - \{\text{CONTENT}\})$,

if the values corresponding to $bvar$, for all occurrences of $bvar$ in a , are constants or C-variables, then

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $bvar$ in a by b ;

otherwise $v_R(e)$ is not applicable to a .

- When $e = (bvar, b) \in BVAR \times \{\text{CONTENT}\}$,

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $bvar$ in a by b .

2. Constant Restriction

- When $e = (cvar, c) \in CVAR \times C$,

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $cvar$ in a by “ c ”.

3. Attribute-Value-Pair Restriction

- When $e = (pvar, (bvar, cvar)) \in PVAR \times (BVAR \times CVAR)$,

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $pvar$ in a by the pair $bvar=cvar$.

- When $e = (pvar, (b, a'')) \in PVAR \times (\{\text{CONTENT}\} \times \mathcal{A}_R)$,

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $pvar$ in a by the pair $b=a''$.

4. P-variable Expansion

- When $e = (pvar, \{pvar_1, \dots, pvar_n\}) \in PVAR \times 2^{PVAR}$,

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $pvar$ in a by the sequence of $pvar_1, pvar_2, \dots$, and $pvar_n$ separated by whitespaces.

5. Tag Type Restriction

- When $e = (tvar, t) \in TVAR \times T$,

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $tvar$ in a by t .

6. Term Restriction

- When $e = (tvar, a'') \in TVAR \times \mathcal{A}_R$,

if all $tvar$ terms² nested in a (in any level) only occur in atomic form, then

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing all occurrences of $tvar$ terms, written as “ $\langle tvar \rangle$ ”, in a by the term a'' ;

otherwise $v_R(e)$ is not applicable to a .

- When $e = (tvar, h) \in TVAR \times \mathcal{H}_R$,

if no $tvar$ term nested in a (in any level) occurs in atomic form, then

$$v_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by replacing each $tvar$ term nested in a by h and the EXP attribute in h by attributes of that $tvar$ term;

otherwise $v_R(e)$ is not applicable to a .

- When $e = (tvar, \varepsilon) \in TVAR \times \{\varepsilon\}$,

if all $tvar$ terms nested in a (in any level) take the form $\langle tvar \rangle \text{CONTENT} \langle /tvar \rangle$, where $\text{CONTENT} \in C \cup \mathcal{A}_R$, then

² $tvar$ terms have the form $\langle tvar \text{ attr}_1 \dots \text{attr}_n \rangle \langle /tvar \rangle$, where $n \geq 0$.

$$\nu_R(e)(a) = a' \in \mathcal{A}_R,$$

where a' is obtained by removing all occurrences of $\langle tvar \rangle$ and $\langle /tvar \rangle$ from a ;

otherwise $\nu_R(e)$ is not applicable to a .

□

Definition 12 (\mathcal{S}_R , the set of specializations, and the mapping μ_R)

Let $\mathcal{S}_R = (\mathcal{C}_R)^*$, the set of all sequences on \mathcal{C}_R . Based on ν_R , the mapping $\mu_R: \mathcal{S}_R \rightarrow \text{partial_map}(\mathcal{A}_R)$ is:

$$\mu_R(\lambda)(a) = a, \text{ where } \lambda \text{ denotes the null sequence,}$$

$$\mu_R(e \cdot s)(a) = \mu_R(s)(\nu_R(e)(a)), \text{ where } e \in \mathcal{C}_R, s \in \mathcal{S}_R \text{ and } a \in \mathcal{A}_R.$$

Note that $\mu_R(s)(a)$ is defined only if all elements in s are successively applicable to a . □

Definition 13 (Specialization system for RDF terms)

The specialization system for RDF terms is $\Gamma_R = \langle \mathcal{A}_R, \mathcal{C}_R, \mathcal{S}_R, \mu_R \rangle$. □

Proposition 1 (Axiomatic requirements of the specialization system for RDF terms)

The specialization system for RDF terms in Definition 13 satisfies the three requirements of specialization systems, i.e.,:

1. $\forall s_1, s_2 \in \mathcal{S}_R, \exists s \in \mathcal{S}_R: \mu_R(s) = \mu_R(s_1) \circ \mu_R(s_2)$,
2. $\exists s \in \mathcal{S}_R, \forall a \in \mathcal{A}_R: \mu_R(s)(a) = a$,
3. $\mathcal{C}_R \subset \mathcal{A}_R$. □

Proof

1. In order to verify that Γ_R satisfies the first requirement, let $s_1 = e_1 e_2 \dots e_n$ and $s_2 = e_1' e_2' \dots e_m'$. The definition of μ_R shows that there exists $s = e_1' e_2' \dots e_m' e_1 e_2 \dots e_n$ such that $\mu_R(s) = \mu_R(s_1) \circ \mu_R(s_2)$.
2. Obviously, $\mu_R(\lambda)(a) = a$, for each $a \in \mathcal{A}_R$, where λ is the null sequence.

3. Definition 9 has already states that \mathcal{C}_R is a subset of \mathcal{A}_R . ■

3.3. Specialization System for RDF Elements

The specialization system for *RDF elements* will be constructed by means of the specialization system for *RDF terms*, defined in Sub-section 3.2.

Definition 14 (\mathcal{A} , the set of RDF elements)

Let \mathcal{A} be the set of all formulae in form of $\langle \text{RDF} : \text{RDF} \rangle t_1 t_2 \dots t_n \langle / \text{RDF} : \text{RDF} \rangle$ where t_i is an RDF term in \mathcal{A}_R . Elements in \mathcal{A} are called *RDF elements*. □

Similar to RDF terms, RDF elements can also be represented equivalently by RDF trees. Denote the RDF tree of the element a in \mathcal{A} by $\text{tree}(a)$.

Definition 15 (\mathcal{G} , the set of ground RDF elements)

\mathcal{G} is that subset of \mathcal{A} which consists of all ground objects in \mathcal{A} . □

Definition 16 (\mathcal{S} , the set of specializations, and the mapping μ)

Let $\mathcal{S} = \mathcal{S}_R$. The mapping $\mu: \mathcal{S} \rightarrow \text{partial_map}(\mathcal{A})$ is defined by:

$$\begin{aligned} \mu(s)(\langle \text{RDF} : \text{RDF} \rangle t_1 t_2 \dots t_n \langle / \text{RDF} : \text{RDF} \rangle) \\ = \langle \text{RDF} : \text{RDF} \rangle \mu_R(s)(t_1) \mu_R(s)(t_2) \dots \mu_R(s)(t_n) \langle / \text{RDF} : \text{RDF} \rangle, \end{aligned}$$

where $s \in \mathcal{S}$ and $t_i \in \mathcal{A}_R$. □

Definition 17 (Specialization system for RDF elements)

The specialization system for RDF elements is $\Gamma = \langle \mathcal{A}, \mathcal{G}, \mathcal{S}, \mu \rangle$. Elements of \mathcal{A} , \mathcal{G} and \mathcal{S} are called *atoms*, *ground atoms* and *specializations*, respectively. □

Proposition 2 (Axiomatic requirements of the specialization system for RDF elements)

The specialization system for the *RDF* elements in Definition 17 satisfies the three requirements of specialization systems, i.e.,:

1. $\forall s_1, s_2 \in \mathcal{S} \exists s \in \mathcal{S}: \mu(s) = \mu(s_1) \circ \mu(s_2)$,
2. $\exists s \in \mathcal{S} \forall a \in \mathcal{A}: \mu(s)(a) = a$,
3. $\mathcal{G} \subset \mathcal{A} \quad \square$

Proof Obvious from the definition of Γ_R . ■

3.4. RDF Declarative Program

After the specialization system for *RDF* elements is defined (cf. Definition 17), the definitions of *RDF* definite clause, *RDF* declarative program and the declarative semantics of an *RDF* declarative program are obtained directly from Definitions 3, 4 and 5, respectively.

```

P = { c1 : a1 ← .
      c2 : a2 ← .
      c3 : a3' ← a3.
      c4 : a4' ← a4. },
where
a1 = <RDF:RDF>
      <RDF:Description RDF:HREF="http://www.example.com/resource1/">
        <DC:Title>Browsing the Internet</DC:Title>
        <DC:Creator>Scott Ring</DC:Creator>
        <DC:Subject>Internet</DC:Subject>
        <DC:Subject>WWW</DC:Subject>
        <DC:Subject>HTML</DC:Subject>
      </RDF:Description>
    </RDF:RDF>,
a2 = <RDF:RDF>
      <RDF:Description RDF:HREF="http://www.example.com/resource2/">
        <DC:Title>Searching on the Web</DC:Title>
        <DC:Creator>John Smith</DC:Creator>
        <DC:Subject>Internet</DC:Subject>
        <DC:Subject>WWW</DC:Subject>
        <DC:Subject>Resource Discovery</DC:Subject>
      </RDF:Description>
    </RDF:RDF>,
a3' = <RDF:RDF>
      <TVAR:A>
        <DC:Creator>
          <RDF:Description>
            <BIB:Name>John Smith</BIB:Name>
            <BIB:Email>john@ait.ac.th</BIB:Email>
            <BIB:Affiliation>Asian Institute of Technology</BIB:Affiliation>
          </RDF:Description>
        </DC:Creator>
      </TVAR:A>
    </RDF:RDF>,
a3 = <RDF:RDF>
      <TVAR:A>
        <DC:Creator>John Smith</DC:Creator>
      </TVAR:A>
    </RDF:RDF>,
a4' = <RDF:RDF>
      <TVAR:B>AIT</TVAR:B>
    </RDF:RDF>,
a4 = <RDF:RDF>
      <TVAR:B>Asian Institute of Technology</TVAR:B>
    </RDF:RDF>.

```

Figure 2. Declarative Program *P*.

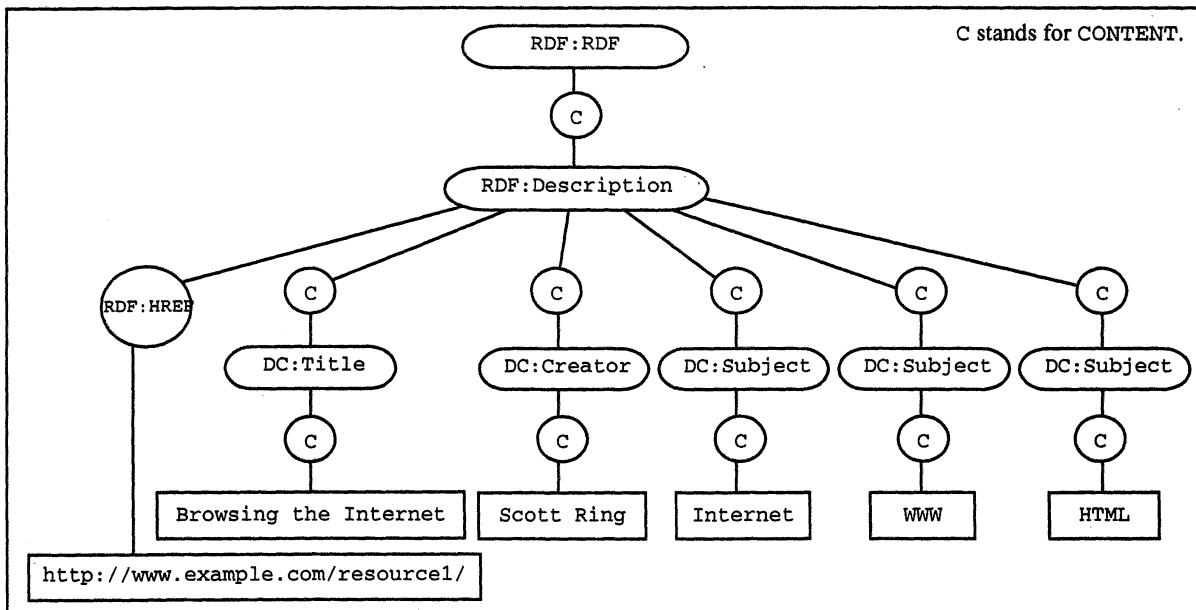


Figure 3. $tree(a_1)$, the RDF tree of atom a_1 .

4 Application to Web-based Resource Discovery Problems: An Example

A Web resource described by *RDF* metadata can be represented directly as a ground *RDF* element in \mathcal{G} . In addition to these representations, rules can be used to explicitly define (possibly complex) relationships between *RDF* elements. A rule is simply written as an *RDF* definite clause C , where $head(C)$ is true, if all atoms in $body(C)$ are true, whence a collection of Web resources can be specified by an *RDF* declarative program P which comprises unit clauses, representing selected Web resources, and non-unit clauses, representing rules. The declarative meaning of P then yields the set of ground atoms each of which represents one Web resource in the specified collection. A query, which expresses a user's information need (possibly implicitly) is represented by an atom in \mathcal{A} , and the result of this query will be the set of Web resources which can be specialized from the query and is contained in the declarative meaning of P .

Based on these modeling concepts, a resource discovery problem is formulated as an intersection problem [2] between the set of *RDF* elements that describe Web resources and the set of *RDF* elements that a query represents.

Example: Let a declarative program P on Γ , which specifies a collection of Web resources, be defined in Figure 2. Unit clauses c_1 and c_2 represent *RDF* elements describing Web resources, non-unit clause c_3 defines a rule which provides additional bibliographic information of John Smith and non-unit clause c_4 gives the abbreviated name of Asian Institute of Technology. In addition to these rules, which are intended to be simple for illustration, complex rules can also be defined when one has to deal with more complicated relationships between *RDF* elements. *RDF* trees representing those atoms in P are depicted in Figures 3 - 6.

A query which finds **only** documents whose authors' affiliations are AIT and contain information about Internet can be formulated as an atom $q \in \mathcal{A}$, i.e.,:

```

q = <RDF:RDF>
  <RDF:Description PVAR:C>
    <DC:Creator>
      <RDF:Description PVAR:D>
        <BIB:Affiliation>
          AIT
        </BIB:Affiliation>
      </RDF:Description>
    </DC:Creator>
    <DC:Subject>
      Internet
    </DC:Creator>
  </RDF:Description>
</RDF:RDF>.

```

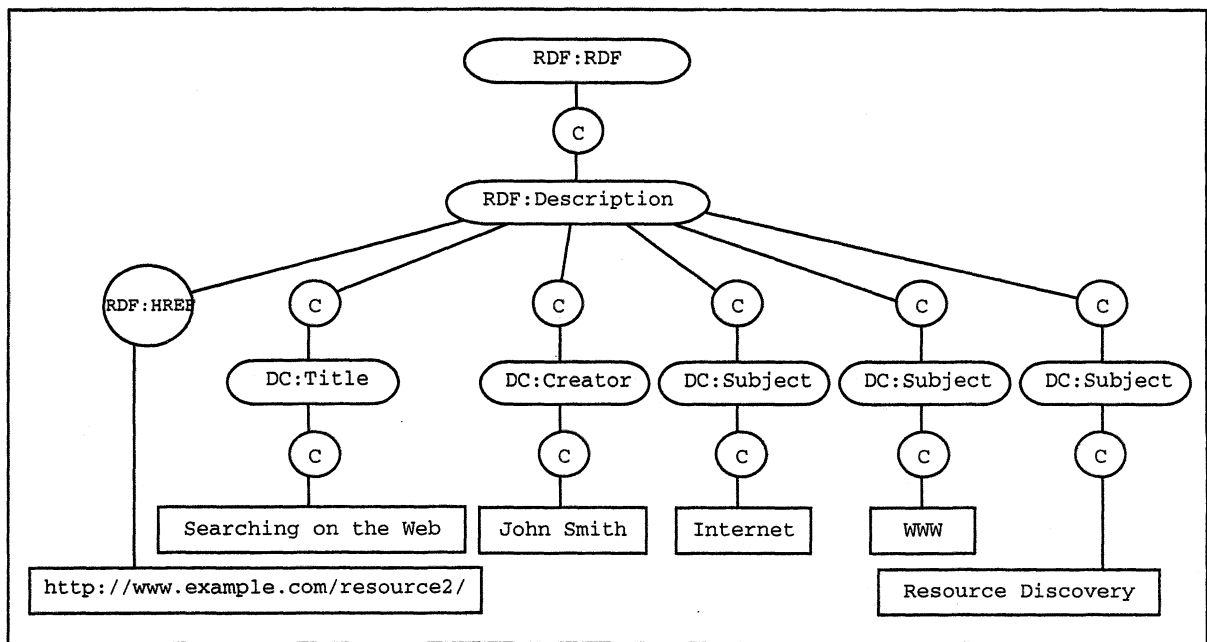


Figure 4. $tree(a_2)$, the RDF tree of atom a_2 .

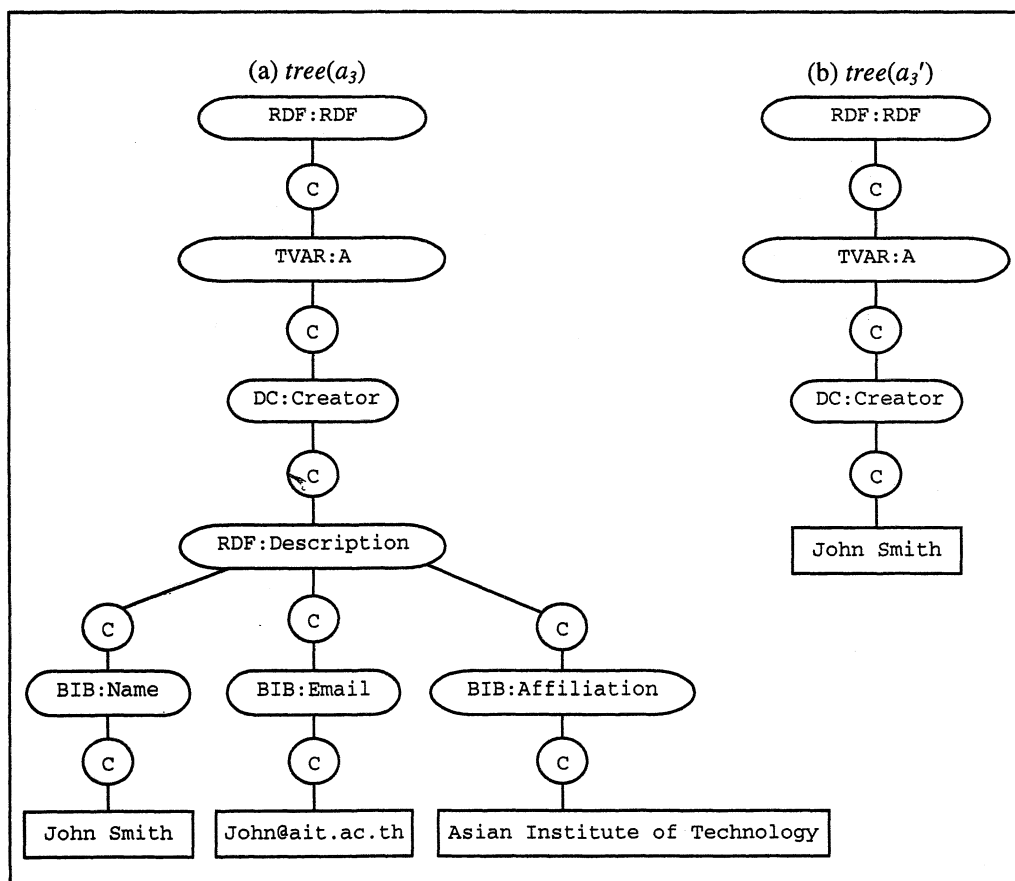


Figure 5. $tree(a_3)$ and $tree(a_3')$, the RDF trees of atoms a_3 and a_3' .

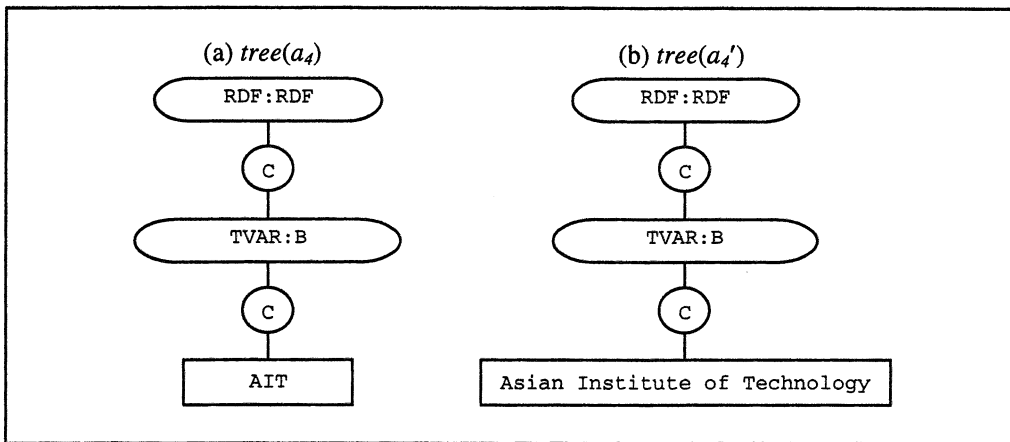


Figure 6. $tree(a_4)$ and $tree(a_4')$, the RDF trees of atoms a_4 and a_4' .

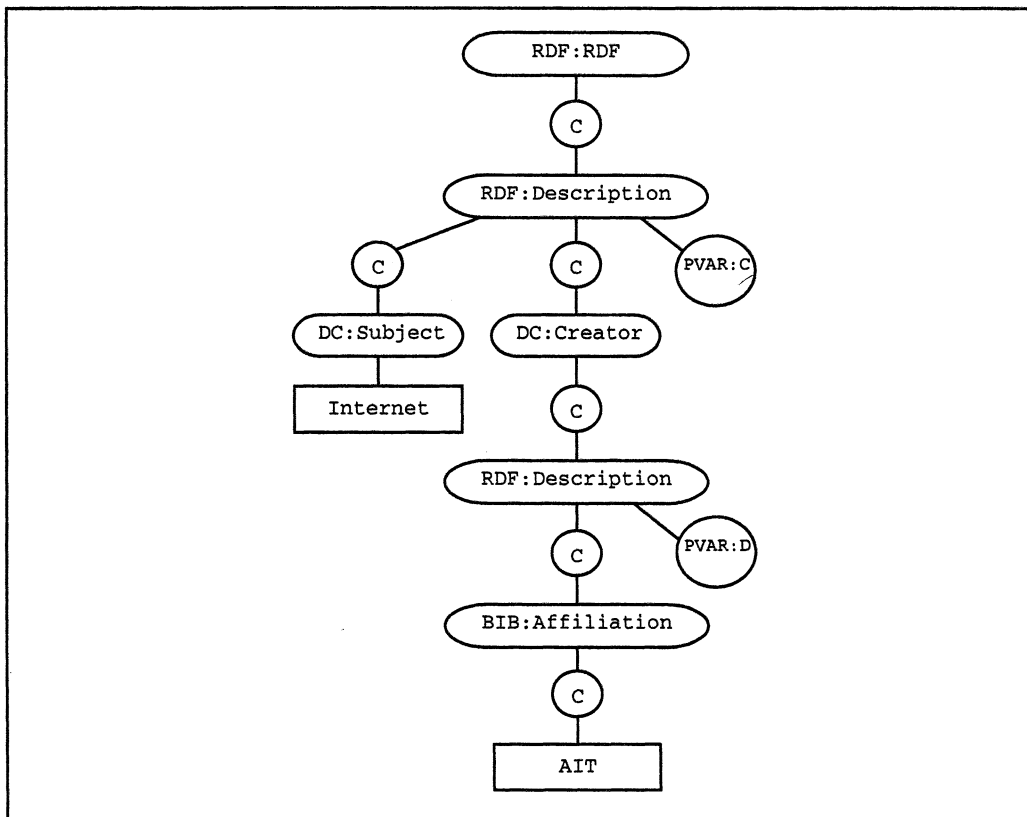


Figure 7. Query atom q and $tree(q)$, the RDF tree of atom q .

Figure 7 illustrates the RDF tree of the query atom q . In processing such a query, first, define atoms a_2' and $a_2'' \in \mathcal{G}$ as in Figure 8. Recalling the definition of T_P (cf. Definition 4), one obtains:

$$\begin{aligned}
 T_P^1(\emptyset) &= \{a_1, a_2\}, \\
 T_P^2(\emptyset) &= \{a_1, a_2, a_2'\}, \\
 T_P^3(\emptyset) &= \{a_1, a_2, a_2', a_2''\}, \text{ and} \\
 T_P^4(\emptyset) &= T_P^3(\emptyset).
 \end{aligned}$$

```

 $a_2'$  = <RDF:RDF>
  <RDF:Description RDF:HREF="http://www.example.com/resource2/">
    <DC:Title>Searching on the Web</DC:Title>
    <DC:Creator>
      <RDF:Description>
        <BIB:Name>John Smith</BIB:Name>
        <BIB:Email>john@ait.ac.th</BIB:Email>
        <BIB:Affiliation>Asian Institute of Technology</BIB:Affiliation>
      </RDF:Description>
    </DC:Creator>
    <DC:Subject>Internet</DC:Subject>
    <DC:Subject>WWW</DC:Subject>
    <DC:Subject>Resource Discovery</DC:Subject>
  </RDF:Description>
</RDF:RDF>

 $a_2''$  = <RDF:RDF>
  <RDF:Description RDF:HREF="http://www.example.com/resource2/">
    <DC:Title>Searching on the Web</DC:Title>
    <DC:Creator>
      <RDF:Description>
        <BIB:Name>John Smith</BIB:Name>
        <BIB:Email>john@ait.ac.th</BIB:Email>
        <BIB:Affiliation>AIT</BIB:Affiliation>
      </RDF:Description>
    </DC:Creator>
    <DC:Subject>Internet</DC:Subject>
    <DC:Subject>WWW</DC:Subject>
    <DC:Subject>Resource Discovery</DC:Subject>
  </RDF:Description>
</RDF:RDF>

```

Figure 8. Atoms a_2' and a_2'' .

Hence, the declarative meaning of P (cf. Definition 5), $\mathcal{M}(P)$, includes atoms a_1 , a_2 , a_2' and a_2'' . Since query atom q can be specialized to atom a_2'' , i.e., there exists $\theta \in \mathcal{S}$ such that $q\theta = a_2''$, the resource whose URI is <http://www.example.com/resource2/> will be retrieved.

5 Conclusions

The RDF is a kind of knowledge/information representation which has no inference mechanism; hence computation with its elements is very limited. Apart from computation by pattern matching, other means of computation is difficult to devise under the present RDF form. This paper has proposed and developed a theoretical foundation upon which reasoning with RDF elements can be carried out. Consequently, complicated computations with RDF elements become possible. Complex queries, like those of SQL in databases, about contents of the resource described by RDF elements can be formulated and handled. This is part of the future research work.

References

1. Akama, K., 1993. Declarative Semantics of Logic Programs on Parameterized Representation Systems. *Advances in Software Science and Technology*, 5:45-63.
2. Akama, K., Kawaguchi, Y. and Miyamoto, E., 1998. Solving intersection problems using equivalent transformation (in Japanese). *Journal of the Japanese Society of Artificial Intelligence* (submitted).
3. Akama, K. and Wuwongse, V., 1997. Equivalent Transformation for Constrained Declarative Programs on String Domains. *Technical Report, Computer Science and Information Management Program, Asian Institute of Technology, Bangkok, Thailand*.
4. Bray, T., Paoli, J. and Sperberg-McQueen, C.M., "Extensible Markup Language (XML) 1.0", Feb 1998. <http://www.w3.org/TR/REC-xml/>.
5. Lassila, O. and Swick, R.R., "Resource Description Framework (RDF) Model and Syntax", 1998. <http://www.w3.org/TR/WD-rdf-syntax/>

A Graph-based Method for Automatic Generation of Multilingual Keyword Clusters and Its Applications

Akiko AIZAWA, Noriko KANDO and Kyo KAGEURA

National Center for Science Information Systems
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-8640 Japan
Tel:+81-3-3942-6994,6968 ; Fax:+81-3-5395-7064
E-Mail: akiko@rd.nacsis.ac.jp

Abstract We are investigating an approach to automatic generation of Japanese-English bilingual keyword clusters using the keyword lists assigned to academic papers by the authors. The bilingual clusters generated by our graph-based method contain keywords with similar meanings from both languages and could be valuable linguistic resources in various information retrieval (IR) applications. In this paper, we first present an overview of our clustering method and then show several experimental results to evaluate the generated clusters. We also discuss the limitation and possible extensions of the current implementation.

keywords cross-lingual information retrieval, automatic extraction of thesaurus, bilingual corpus, graph theory, academic paper database

1 Introduction

The explosive growth of online documents has increased the need for IR systems that cross language boundaries. For instance, since the first workshop on cross-lingual information retrieval (CLIR) in 1996 (Grefenstette, Smeaton & Sheridan, 1996), the topic has been one of the most actively pursued in IR research field. One especially interesting and important research to this direction is the *automatic* generation of domain-dependent multilingual thesaurus; this not only help searchers of multilingual scientific databases but also are useful in monolingual search since technical terms are often imported either in their original forms or as acronyms, transliterated, and then translated (Kando 1997).

Presently, we are investigating an approach to automatic generation of Japanese-English bilingual keyword clusters using the

keyword lists assigned to academic papers by the authors (Aizawa & Kageura, 1998). The bilingual clusters generated by our graph-based method contains keywords with similar meanings from both languages and could be valuable linguistic resources in various IR applications.

The basic idea of our clustering strategy is that we apply graph theoretic method instead of statistical ones which seem dominant in corpus-based approaches; the bilingual keyword pairs constitute a tangled graph of Japanese and English keywords; as such, the clustering problem can be regarded as a problem of partitioning the original keyword graph by eliminating wrongly generated links from the graph; then, the problem can be transformed into the *minimum cut problem* in the graph theory.

Applying graph theoretic method has several advantages. First, low-frequency key-

words can be treated properly by utilizing topological features of the graph. Second, the clusters contain not only Japanese-English pairs, but also Japanese-Japanese and English-English pairs. Thus, they are usable for not only CLIR but also query expansion in monolingual IR. Third, the whole process can be achieved with reasonable computational cost.

The keyword data in view of bilingual corpus also has several advantages. In conventional corpus-based approaches for CLIR (Dunning & Davis, 1993; Landauer & Littman, 1990; Carbonell, 1997), the lack of readily available parallel corpora has been a bottleneck. On the other hand, our corpus is available for a great many subject domains, with rich variations that may not be listed in the standard dictionaries but are nevertheless meaningful and useful in IR. Another advantage of our keyword corpus is that keyword pairs can be easily extracted without expensive natural language processing for segmentation and alignment.

In the following, we first briefly summarize the outline of our clustering method and then report several experimental results to evaluate the generated clusters; comparison with standard dictionaries, analyzing errors in the clustering, and the performance with bilingual and monolingual IR. We also discuss the limitation and possible extensions of the current implementation.

2 Overview of Multilingual Keyword Cluster Generation Procedure

Our procedure for generating multilingual keyword clusters is composed of the following stages: (1)extraction of Japanese-English keyword pairs from basic corpus, (2)simple normalization, (3)generation of initial keyword graph, (4)screening obvious non-errors, (5)detection of possible correspondence errors, (6)detection of possible homonymous words, (7)partitioning keyword clusters, and

(8)output of final clustering results. Each stage is described briefly in the following.

2.1 Extraction of Japanese-English Keyword Pairs

The basic data used in the current study is the Japanese and English keywords assigned by the authors to their papers, extracted from the NACSIS Academic Conference Database (NACSIS, 1997). We selected 28,122 papers related to the field of computer science. Of the papers selected, 26,060 (about 93 %) have the same number of Japanese and English keywords. An example is :

Japanese: 多言語検索 / 用語クラスタ / グラフ理論 / NACSIS データベース

English: cross-lingual information retrieval / keyword cluster / graph theory / NACSIS Database.

Since the Japanese and English keyword pairs generally maintain good one-to-one correspondences, we mechanically extract total 112,364 Japanese and English keyword pairs (60,186 different ones) and use them as a basic bilingual keyword corpus.

Table 1 shows some examples of the extracted keyword pairs. The most general English translation for each Japanese keyword is marked with '*'.

2.2 Simple Normalization

After the extraction of bilingual keyword pairs, simple normalization is applied to deal with notation variation problem such as *cross-lingual* and *Cross Lingual*. Also, acronyms are detected and marked so that they can be tested for homonyms at later stage.

2.3 Generation of Initial Keyword Graph

The initial graph expression of a bilingual keyword corpus is easily derived by representing Japanese and English keywords as nodes and their translation pairs as links. The frequency of the keyword pair appeared in the

Table 1: Example of Japanese-English keyword correspondences extracted from the database.

Japanese keywords	English keywords	frequency
キーワード	information retrieval	1
キーワード	keyword	39*
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6*
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retrieval	1
情報検索	information retrieval	320*
情報検索	information search	5
情報収集	information gathering	6*
情報収集	information retrieval	1
文献検索	bibliographic search	1*
文献検索	document retrieval	11
文書検索	document retrieval	19*
文書検索	text retrieval	1

corpus is expressed as the capacity of the corresponding links. Figure 1 shows the initial keyword graph generated from the keyword pairs shown in Table 1.

The global keyword graph is composed of numbers of disjoint sub-graphs, which we define as *bilingual keyword clusters*. In case of the above example, the whole nodes belong to the same keyword cluster at the initial stage and thus are considered to have similar meanings.

2.4 Screening Obvious Non-Errors

This stage currently includes (1) recognition of Japanese and English pairs with identical notations, (2) checking keyword pairs with sufficient frequencies, and (3) detecting *minor examples*. The last case applies when a keyword pair stands for the unique translation of either of the Japanese or English counterpart. Links which satisfy the above conditions are considered to be correct and maintained automatically to reduce the computation cost

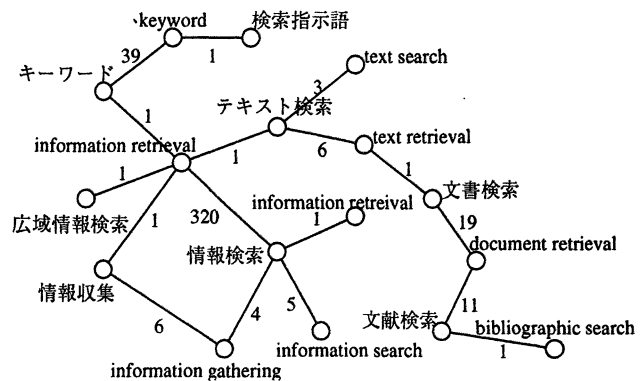


Figure 1: Initial keyword graph generated from keyword pairs in Table 1.

at later stages.

2.5 Detection of Possible Correspondence Errors

The initial keyword graph constructed from 60,186 different translation pairs contains a huge keyword cluster with as many as 20,659 nodes (about 34%) in it. Since our subject is restricted to the specific academic field *computer science*, the major cause for this is the existence of improper translation pairs which connects otherwise disjoint keyword clusters into one (This is in contrast to general cases where the homonyms are much more common). A typical example found in Figure 1 is the link <キーワード (*keyword*), information retrieval >.

The detection algorithm employed in our procedure is based on a simple principle that *a set of links which decompose a connected keyword cluster into disjoint sub-clusters when they are removed from the original cluster are the candidates of improper translations*. In the conventional graph theory, such a link set is called an *edge cut* and the edge cut with the minimal total capacity among all the edge cuts obtained for a graph is called a *minimum edge cut*. Minimum edge cut problem is one of the most principal prob-

lems in the graph theory and there exist number of algorithms which guarantee sufficient performance for our purpose.

2.6 Detection of Possible Homonyms

Though homonyms do not seem to occur so frequently in a specific scientific domain, we still have observed several cases such as < ATM (*Asynchronous Transfer Mode*) > and < ATM (*Automatic Teller Machine*) >.

The detection of possibly homonymous keywords can also be done utilizing the topological feature of the keyword cluster. It can be assumed that *homonymous nodes are the ones which decompose the cluster when the node and all the edges starting from the node are removed*. Thus, the problem is transformed again to the well-known node cut problem of the graph theory. Since most of the homonyms we have observed are acronyms, we presently consider only keywords composed of English capital characters and symbols as candidates as node cuts.

2.7 Partitioning Keyword Clusters

The minimum edge cut of a keyword cluster does not always represent imprecise translation. Removing correct pairs inevitably causes oversplitting, i.e. generating more than one clusters with similar meanings. On the other hand, the distinction between corresponding keyword pairs and associated but not corresponding ones depends on the application and is difficult even for human experts. For example, the keyword pair < テキスト検索 (*text retrieval*), information retrieval > may be improper in view of strict terminological definition but not necessarily be incorrect for searchers of academic paper databases.

Our current implementation employs a simple stopping criteria: partitioning occurs only when (1) the total capacity of the minimum edge cut is equal or less than N_α , and also (2) each of the newly generated clusters contains at least one nodes with greater than

N_β frequencies. We presently set $N_\alpha = N_\beta (= N)$ and use the same value for all the clusters.

Once candidates for improper translations are obtained, partitioning is done automatically by removing the links on the original graph expression. Homonyms can be similarly processed by splitting the nodes into different clusters. The detection and deletion stages described so far is applied for each keyword cluster recursively until no more pairs can be removed.

2.8 Final Clustering Results

Figure 2 shows the result of the partitioning of the keyword cluster given in Figure 1.

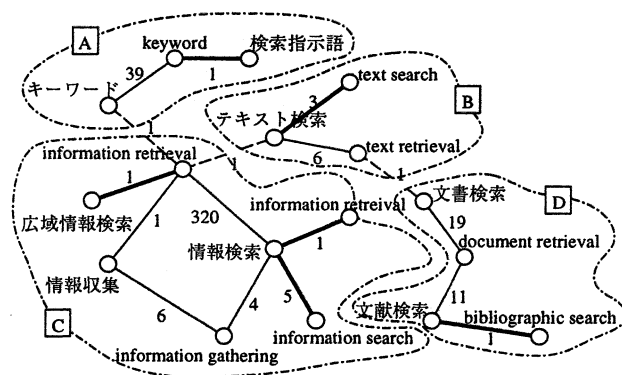


Figure 2: Example of partition of keyword cluster.

As a result of the detection of correspondence errors, three keyword pairs < キーワード (*keyword*), information retrieval >, < テキスト検索 (*text retrieval*), information retrieval >, < 文書検索 (*document retrieval*), text retrieval >, are removed, and four clusters A, B, C, and D are newly created. The bold lines shows that the links are marked as unremovable at screening stage. This follows that such pairs as < 情報検索 (*information retrieval*), information retrieval > (spelling error), < 検索指示語 (*keyword*), keyword > (rare case), and < 広域情報検索 (*wide-area information retrieval*), information retrieval >

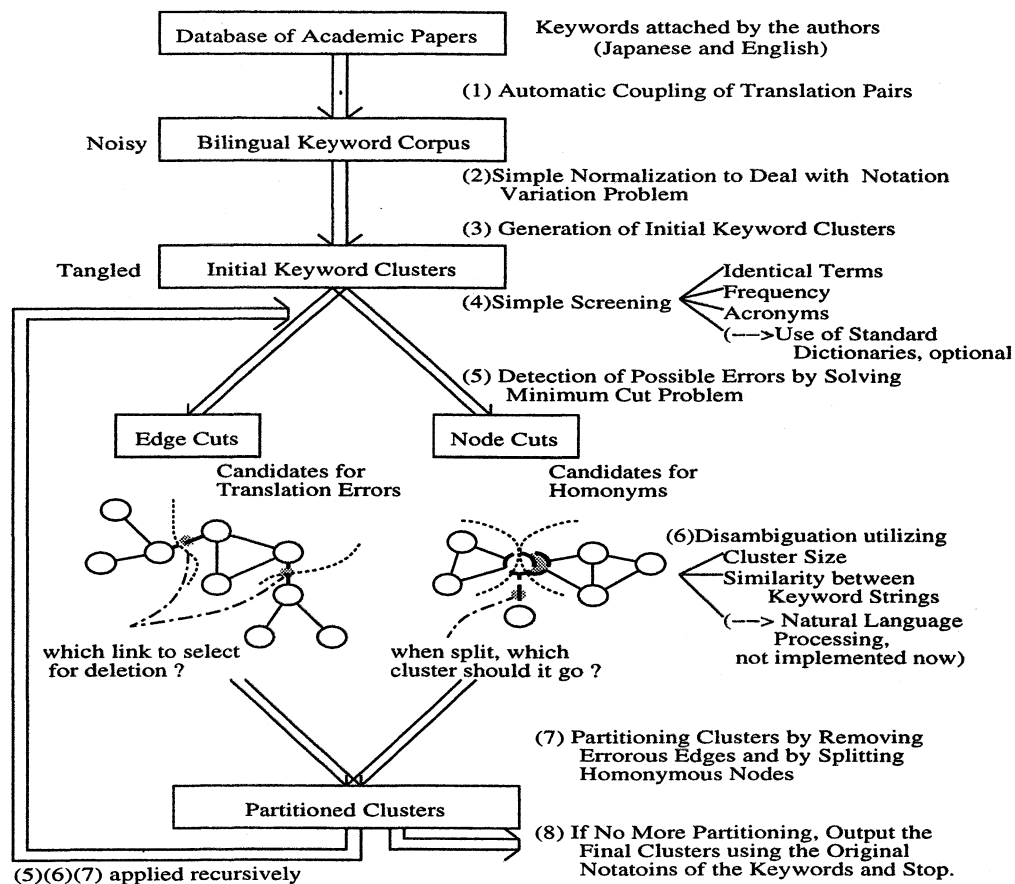


Figure 3: Overview of the proposed keyword clustering procedure.

(related but not equivalent pair) are maintained even after the partitioning.

Applying the method to the whole corpus with $N = 5$, 1,469 of the total 60,186 keyword pairs were deleted, generating the total 27,918 keyword clusters. The number of keyword pairs included in the biggest cluster was reduced to 159 from 20,659 of the initial graph. The overall procedure described in this section is illustrated in Figure 3.

3 Experimental Results

3.1 Comparison and integration with standard dictionaries

In Table 2, the keyword corpus we use in our study is compared with total 22,690 different term pairs extracted from four dictionar-

ies and handbooks of the field of computer science (Aiso, 1993; Japan Society for Artificial Intelligence 1990; Ralston, 1983; Shapiro, 1987).

The comparison shows that the number of common elements between the corpus data and the standard technical dictionaries is relatively small, i.e. many of the keyword pairs assigned by the authors are not listed in the standardized technical dictionaries. This may partly be because the keyword data contains some noises such as spelling errors or notation variations¹. The noises themselves are useful in IR task since they may also be

¹The number of common elements are increased to 3,074 from 2,066 after simple normalization to deal with notation variation.

Table 2: Comparison of the technical dictionaries and the bilingual keyword data.

	dictionary terms	corpus pairs	common for both
Japanese words	20,636	37,170	3,966
English words	19,562	49,918	2,814
different translation pairs	22,690	60,186	2,066
average number of translations per word(Jpn)	1.10	1.62	—
average number of translations per word(Eng)	1.16	1.21	—
maximum number of translations per word(Jpn)	7	86	—
maximum number of translations per word(Eng)	6	29	—
number of English acronyms (Jpn)	844	1,007	212
number of English acronyms (Eng)	451	1,233	114
identical Japanese and English pairs	57	1,336	18

found in queries of databases.

Another important point shown in the table is that the Japanese keywords often contains English acronyms in both standard dictionaries and in keyword corpus. The fact supports our assumption that bilingual keyword clusters should be valuable not only for CLIR but also in Japanese monolingual search.

Table 3 shows how the ratio of common pairs between the dictionaries and the corpus changes against the frequencies they appear in the corpus. We can observe that the more frequent a keyword pair appears in the corpus, the higher is the probability it is also referred in the standard dictionaries. For example, among 68 pairs with more than 100 frequencies, only 10 are *not* referred in the standard dictionaries. They are mostly acronyms or terms representing relatively new technologies: i.e. { CAI, CASE, CSCW, LOTOS, OSI, WWW, agent, multi media, genetic algorithm, reuse }. The average frequency of the common pairs in the corpus is 5.9 before normalization and 11.1 after normalization.

In conclusion, it can be expected that the keyword corpus reflects particular views and concepts of the authors, which can not be covered with the standard dictionaries. Upon

Table 3: The ratio of common terms against the frequency.

frequency more than N	num. of keyword pairs	num. of common pairs	ratio
1	51062	3074	0.06
2	10990	1999	0.18
5	2916	1062	0.36
10	1216	606	0.50
20	538	326	0.61
50	176	127	0.72
100	68	58	0.85
200	24	24	1.00

a extraction of basic keyword corpus, words from standard dictionaries can easily be integrated (with frequencies set to the infinite) to introduce more generality in the original corpus data. The effect of such integration is examined in later through IR evaluations.

3.2 Analyzing clustering result

The noise in the keyword corpus can be categorized into either of the following groups (the examples shown for each group are the English correspondings to the Japanese term “情報検索 (*information retrieval*)”):

Table 4: Errors detected at the first iteration of partitioning.

total errors detected	951
(1) spelling error	0
(2) expressional variations	0
(3) related keywords	598
(4) obvious errors	326
(5) correct	27

- (1) spelling errors
example: information retrieval
- (2) expressional variations
example: information retrieving
- (3) related keywords
example: information seeking
- (4) obvious errors
example: keyword

When generating keyword clusters, it is advantageous to maintain spelling errors and expressional variations in view of recall, while it is important to eliminate obvious errors in view of precision. The treatment of related keywords should depend on the database and also the context of the search.

We analyze manually the errors detected at the first iteration of the cluster partitioning, i.e. the minimum cut links with the capacity equal to 1. The result is summarized in Table 4. Of the total 951 pairs detected as errors, 326 are obvious errors, 598 are related but not exactly corresponding, and 27 are detected wrongly though they actually are the correct ones. It is remarkable that regardless of the considerable number of spelling errors and expressional variations included in the database, non of them are detected as errors. This may be because such errors or variations occur most likely with low frequency and seldom are repeated in different clusters.

Among 598 related pairs, 136 have hierarchical relationship while the rest are the ones simply correlated. Among 27 detection er-

rors, 10 are caused by homonymous keywords that are not acronyms, 8 by mis-processed acronyms, and 9 by minor examples that errors actually occur more frequently than correct pairs in the corpus.

The result shows that the major improvement can be expected by refining conditions for links elimination after minimum cuts detection stage. Our present implementation use only graph-topological conditions, N_α and N_β , common for all the clusters. Better results may be obtained by changing the value depending on the cluster size or by applying natural language processing. Since the number of candidates are greatly reduced by minimum cuts detection stage, the disambiguation of errors and non-errors can be more time consuming.

In the following, a few examples of the generated keyword clusters are shown where the number in the parenthesis indicates the frequency of the keyword. Example 1 shows the keywords in the biggest cluster with their frequencies ≥ 3 . We can observe the closely related Japanese and English keywords are clustered together.

Example 1: *Frequent keywords in the largest cluster.*

Japanese: 並列処理 (740), 並列 (62), 並列化 (56), 並列計算 (29), 並行処理 (19), 並列性 (10), 並列システム (6), 多重処理 (6), 並行システム (5), マルチプロセス (5), 並列プロセス (4), 並列度 (3), 多重プロセス (3), マルチプロセッシング (3), パラレル処理 (3)

English: parallel processing(672), parallel(74), parallelization(44), concurrent processing (20), parallel computing(18), parallel computation(18), parallelism(14), parallel process(8), multi process(8), concurrent system(7), parallel system(6), parallel processings(6), multi processing(6), multi-processing(5), multiprocess(5), concurrent(5), future(4), parallelize(4), parallel processes(4), parallel operation(4), paralell processing(4)

Example 2 contains two clusters which are originally a single cluster but separated since each of the clusters has sufficient frequency in the target corpus. Again, we show only the keywords with their frequencies ≥ 3 on

account of the limited space.

Example 2 : *Associated clusters divided into two.*

(1) CLUSTER 1

Japanese: 知的 *cai*(118), *ITS*(67), 知的教育システム (31), *ICAI*(11), 知的教授システム (8), 指導方略 (7), 教授戦略 (7), 問題演習 (4), 定式化 (4), 教育戦略 (4), 対象理解 (3), 教育効果 (3)

English: *ITS*(68), *intelligent cai*(65), *intelligent tutoring system*(45), *ICAI*(30), *intelligent educational system*(9), *tutoring strategy*(8), *teaching strategy*(8), *IES*(7), *teaching paradigm*(3), *object understanding*(3), *intelligent tutoring systems*(3)

(2) CLUSTER 2

Japanese: *CAI*(156), 教育支援システム (27), 教育支援 (21), 学習支援 (13), 学習支援システム (8), *cai*システム (7), *CAL*(6), 表層構造 (4), 派生語 (4), 学習支援環境 (4), コンピュータ支援教育 (3)

English: *CAI*(169), *computer assisted instruction*(27), *computer aided instruction*(8), *CAL*(7), *education support system*(6), *derivative*(4), *cai system*(4), *surface structure*(3), *learning support*(3), *education support*(3), *computer assisted learning*(3)

Example 3 shows another but smaller cluster with all the Japanese and English keywords included. It can be observed that minor translation examples are integrated into more frequent cases.

Example 3 : *An output including minor keywords.*

Japanese: *CAD*(246), 設計支援 (63), 計算機援用設計 (14), 計算機支援設計 (9), コンピュータ支援設計 (8), コンピュータ援用設計 (4), コンパイルエラー (4), 設計工学 (3), キヤド (3), 設計エンジニアリング (2), 計算機設計支援 (2), ソフトウェア設計支援 (2), 知能 (1), 自動組立 (1), 三面図理解 (1), 航空機主翼設計 (1), 計算援用設計 (1), 演算器シェア (1), コンピュータデザイン (1)

English: *CAD*(225), *computer aided design*(70), *design support*(36), *design engineering*(6), *software design support*(4), *design assistance*(2), *cad*(2), *support to plan*(1), *support method for design*(1), *software design support*(1), *prulog*(1), *planning aids*(1), *operator resource sharing*(1), *design = support*(1), *design support system*(1), *design support*(1), *design support*(1), *design support*(1), *design assist*(1), *design support*(1), *computer aided design*(1), *computer aided design*(1), *computer aided design*(1), *computer aided design*(1), *compile time errors*(1), *compile error*(1), *computer aided design*(1), *assisted design*(1), *DAD*(1)

3.3 Query Expansion in Cross-lingual Information Retrieval

Lastly, We examine the effectiveness of the generated clusters on the two retrieval tasks:

- (1) CLIR: Japanese queries retrieving documents from an English collection (J-E task), and
- (2) monolingual IR: Japanese queries retrieving documents from a Japanese collection (J-J task).

The search performance is tested against the test version of the NACSIS Test Collection 1 (Kando et al, 1998) for the J-E task E collection, which contains 186,809 documents, and for the J-J task, J collection, 338,668 documents.

We indexed Japanese terms by character (uni-gram), and English terms by word. English terms appeared in Japanese texts were also indexed by word. Queries are submitted as Japanese natural language sentences. They were initially segmented into words using a Japanese morphological analyzer, Chasen v1.5 (Matsumoto, et al., 1997). Words and phrases were then automatically selected as query terms using several patterns defined over part-of-speech tags.

Each query term was translated or expanded using the bilingual keyword clusters reported here. We treated terms in the cluster, containing the query term, as synonyms. In order to examine the effect of standard dictionary integration and also partitioning parameter N , we tested these strategies listed below; K3: keyword clusters obtained with $N = 3$, KD3: keyword and dictionary term clusters obtained with $N = 3$, KD10: keyword and dictionary term clusters obtained with $N = 10$, D3: dictionary terms in KD3, and D10: dictionary terms in KD10. The average numbers of translated or expanded terms per a query term are shown in Table 5.

The search engine is OpenText 6, which can handle both English and Japanese char-

Table 5: Average number of expanded terms per a query.

	K3	KD3	KD10	D3	D10
J-E	6.81	4.48	3.66	0.84	0.77
J-J	10.7	7.01	5.72	1.4	1.3

acters. The documents in the returned set are ranked using OpenText's "RankMode Relevancel".

The retrieval results for the J-E task are summarized in Figure 4 where each plot represents the search effectiveness compared with the *baseline*, which is obtained by applying the original Japanese query terms to the Japanese correspondings of the target English documents.

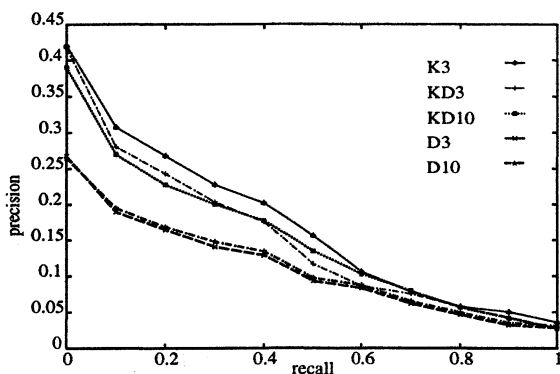


Figure 4: Retrieval results of J-E task.

For the J-E task, the keyword clusters (K3) showed the highest effectiveness, then followed by the keyword and dictionary term clusters (KD3 and KD10). The clusters consisted of dictionary terms only (D3 and D10) achieved less than 65% of the K3. The number of translated terms seems to be an important index to estimate the search effectiveness in the J-E task. No significant difference was found regarding the minimum edge cut levels of 3 and 10 in the keyword and dictionary clusters (KD3, KD10), but some improvement is shown in the dictionary clusters

(D3, D10).

The retrieval results for the J-J task are summarized in Figure 5 where each plot represents the 11 point average search effectiveness over the *baseline*, which in this case is the performance without query expansion.

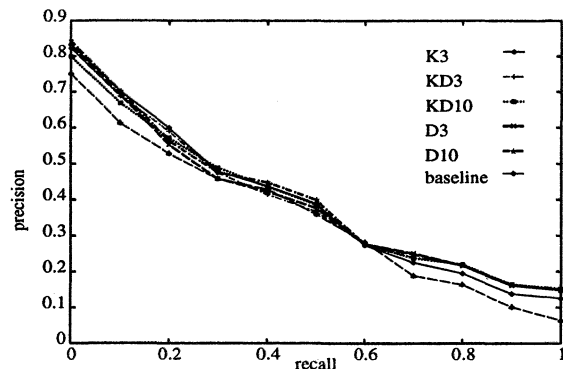


Figure 5: Retrieval results of J-J task.

For the J-J task, all the strategies showed the improvement of search effectiveness of 11.3-13.9% in the average over the baseline. Though the average numbers of expanded terms were significantly small in the dictionary clusters (D3, D10), they showed as same as, or even more improvement of search effectiveness over the baseline. Again, there were no significant difference according the clustering parameter in the keyword and dictionary clusters (KD3 and KD10); Some improvement is observed in the dictionary only clusters (D3 and K10).

Based on the results, we can temporarily conclude that the bilingual clusters generated by our method showed significant improvement of search effectiveness both in CLIR and monolingual IR. In CLIR, the bilingual clusters of author-assigned keywords (K3) were more effective than the ones with dictionary terms, while in monolingual IR, adding dictionary terms to the clusters (KD10) showed improvement. It is also observed that the clustering parameter N affects the performance in some cases, suggesting that fur-

ther improvement can be expected by refining clustering conditions. The IR experiments described here will be explained more in detail (Kando & Aizawa, 1998).

4 Discussions

The keyword clusters generated by our method can also be utilized in automatic indexing such as LSI (latent semantic indexing) to reduce the dimension of the frequency matrix by term clustering. Also, the clusters generated by our graph-based method but using only dictionary terms are utilized in domain map visualization task to assist interactive document retrieval (Aihara & Takasu, 1998).

Though our present implementation use only graph-topological information, we are now looking for the possibility of incorporating natural language processing. For example, only keyword level correspondences are considered so far but given that many keywords are complex, we can expect better performance by utilizing morpheme (or word) level correspondences. This will also give a way to weight the bilingual pairs according to their centrality within a cluster and to control the granularity of the generated cluster.

Acknowledgement

The research reported here is a part of the research project "A Study on Ubiquitous Information System for Utilization of Highly Distributed Information Resources", granted by the Japan Society for the Promotion of Science.

References

- [1] Grefenstette, G., Smeaton, A. and Sheridan, P. (eds.) (1996) *Workshop on Cross-Linguistic Information Retrieval*.
- [2] Kando, N. (1997) *Cross-linguistic scholarly information transfer and database services in Japan*. Presented at the Annual Meeting of the American Society for Information Science.
- [3] Aizawa, A. and Kageura, Kyo (1998) *An Approach to the Automatic Generation of Multilingual Keyword Clusters*. COMPTERM'98.
- [4] Dunning, T. and Davis, M. (1993) *Multilingual information retrieval*. Technical report MCCS-93-252, Computer Research Laboratory, New Mexico State University.
- [5] Landauer, T.K. and Littman, M.L. (1990) *Fully automatic cross-language document retrieval*. In Proceedings of the Sixth Conference on Electronic Text Research, p.31-38.
- [6] Carbonell, J.G., Yang, Y., Frederking, R.E., Brown, R.D., Geng, Y. and Lee, D. (1997) *Translingual information retrieval: a comparative evaluation*. In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97), p.708-714.
- [7] NACSIS (1997) *Introduction to the National Center for Science Information Systems*. NACSIS.
- [8] Aiso, H. (ed.) (1993) *Joho Syori Yogo Daijiten*. Tokyo: Ohm.
- [9] Japan Society for Artificial Intelligence (ed.) (1990) *Jinko Tinou Handobukku*. Tokyo: Ohm.
- [10] Ralston, A. (ed.) (1983) *Encyclopedia of Computer Science and Engineering*. Amsterdam: Van Nostrand Reinhold. [Toujou, A. (trans. ed.) *Compyu-ta Daihyakka*. Tokyo: Asakura. 1987.]
- [11] Shapiro, S (ed.) (1987) *Encyclopedia of Artificial Intelligence*. New York: John Wiley. [Ohsuga, S. (trans. ed.) *Jinko Tinou Daijiten*. Tokyo: Maruzen. 1991.]
- [12] Kando, N., Koyama, T., Oyama, K., Kageura, K., Yoshioka, M., Nozue, T., Matsumura, A. and Kuriyama, K. (1998) *NTCIR: NACSIS Test Collection Project [Poster]*. the 20th Annual BCS-IRSG Colloquium on Information Retrieval Research.
- [13] Matsumoto, Y. et al. (1997) *Japanese Morphological Analyzer Chasen 1.5*. NAIST.
- [14] Kando, N. and Aizawa, A. (1998) *Cross-Lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters* IRAL'98 (submitted).
- [15] Aihara, K. and Takasu, A. (1998) *Domain Visualization Based on Authorized Documents* SCI '98/ISAS '98.

An Intelligent Search for Thai Text in Digital Libraries

Asanee Kawtrakul, Thanussak Thanyasiri, Chavalit Chirarattachan,
Nathavit Buranapraphanont, Preeti Piti-alongkorn, Navapat Khantonthong
Natural Language Processing and Intelligent Information System
Technology Research Laboratory,
Computer Engineering Department Faculty of Engineering, Kasetsart University
Email: [ak,g40tht,g40cvc,g40nab,g40prtp,g40nak]@ku.ac.th

Soontharee Koopairojn
Computer Science Department, Faculty of Science, Kasetsart University
Email: [fscisok@ku.ac.th]

Abstract

One of technologies for next generation of Digital Library is information access technology. Since the direction of future digital library is expected to handling large collections of electronic documents, such as technical articles; journals; thesis; scientific reports, the ability to search for bibliographic records, eg. complex boolean, phrase proximity, is not sufficient. This paper proposes integrating technologies : natural language processing and retrieval processing that enable efficient access to Thai text in Digital Library.

Keywords: Digital Library, Natural language processing, Text Retrieval

AUTHOR INDEX

Adachi, Jun	40
Aizawa, Akiko	94
Anutariya, Chutiporn	82
Baker, Thomas	72
Bright, Myron	46
Chirarattanachan, Chavalit	104
Choi, Han Suk	73
Dartois, Myriam	74
Fullerton, Karen	46
Greenberg, Jane	46
Kageura, Kyo	94
Kando, Noriko	94
Kawtrakul, Asanee	104
Khantonthong, Navapat	104
Kikuta, Masahiro	6
Kita, Katsuichi	64
Koompaiojn, Soontharee	104
Maeda, Akira	74
Maeda, Harumi	64
McClure, Maureen	46
Miller, Eric	71
Nagata, Yoshikatsu	64
Nakao, Shigetaka	74
Nantajeewarawat, Ekawit	82
Nuranaprophanont, Nathavit	104
Ohta, Jun	74
Piti-alongkorn, Preeti	104
Rasmussen, Edie	46
Rowland, Fytton	47
Sakaguichi, Tetsuo	74
Shibata, Youichi	23
Shibayama, Mamoru	64
Stewart, Darin	46
Sugimoto, Shigeo	74
Tabata, Koichi	74
Thanyasiri, Thanussak	104
Weibel Stuart	71
Wuwongse, Vilas	82
Yamamoto, Nobuhito	1

KEYWORD INDEX

academic paper database	94
automatic extraction of thesaurus	94
bilingual corpus	94
computer network	1
computer technology	1
cross-lingual information retrieval	94
DC-based union cataloging system	73
declarative programs	82
digital library	1,40,46,73,104
digital object repository	46
electronic journal	40,47
electronic library service	40
electronic library	40
electronic publishing	47
extensible markup language	23
GEM metadata standard	46
graph theory	94
HTML	6
HTTP	6
K-12 educator	46
microfilm image database	64
microfilm retrieval system	64
multilingual document browsing	74
multilingual texts display and input	74
multimedia	64
natural language processing	104
off-the-shelf WWW browsers	74
online academic journal	73
online journal	40
RDF	71
RDF elements	82
resource discovery problems	82
scholarly communication	47
scholarly journal	47
scholarly publishing	47
SGML	6
text retrieval in multiple scripts	74
URL	6
World Wide Web	6
XLL	23
XML	23,71
XSL	23
Z39.50 standard	73

A digital version of these proceedings
is available at the World Wide Web URL:
<http://www.DL.ulis.ac.jp/DLjournal/>
<http://beethoven.cpe.ku.ac.th/ijwdl98/DLjournal/>